The OBJECTIVE OR NEW-TYPE EXAMINATION

RUCH

— PRESENTED TO —

HE SOUTHWEST CHRISTIAN SEMINARY

Phoenix, Arizona

y Everett Johnson

f Denver, Colo.

Date Sept. 26, 1950

om out the hearts and hands of faithful friends come
e human resources which make our service possible.

THE MARY E. BIVINS LIBRARY
of
THE SOUTHWEST CHRISTIAN SEMINARY
Phoenix, Arizona

Digitized by the Internet Archive
in 2023 with funding from
Kahle/Austin Foundation

https://archive.org/details/objectiveornewty0000gmru

THE OBJECTIVE OR NEW-TYPE EXAMINATION

An Introduction to Educational Measurement

BY

G. M. RUCH

PROFESSOR OF EDUCATION
UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA

THE CINCINNATI BIBLE SEMINARY
LIBRARY

SCOTT, FORESMAN AND COMPANY
CHICAGO ATLANTA NEW YORK

371.271
R899a

Copyright, 1929, by
SCOTT, FORESMAN AND COMPANY

295.1

27203

PREFACE

This volume is intended for two general classes of readers: first, the teacher wishing to make a serious study of the theory and practice of objective examining, and second, the student who is beginning his study of educational measurement. The organization of the book has kept in mind the fact that the general reader will not always be familiar with statistical procedures. For this reason the more technical topics are reserved for the closing chapters.

It is the author's conviction based upon nearly ten years of experience with the construction, administration, and teaching of tests and measurements that the introduction to the theory and practice of educational measurements may best come from an initial study of informal classroom tests. The teacher and the student are already familiar with the usual school examinations. This background provides a basis for the introduction of such concepts as validity, reliability, objectivity, sampling, errors of measurement, etc.

After the basic concepts of educational measurement have been mastered through the avenue of informal objective tests, the student is in a position to evaluate standard tests critically.

This volume has been given the subtitle, *An Introduction to Educational Measurement,* in the belief that such a treatment should precede and introduce formal study of standardized measurements.

Judging by the developments of the past five years, the objective examination has come to stay for an indefinite period. It is almost certain to persist, in one form or another, just as the older forms of examinations will con-

tinue in use, with important modifications. No one method of examining is likely to prove adequate for the varied purposes of the teacher. The future will see the traditional examination, the new-type test, and the standard test exist side by side. We may expect continual efforts at correlation and delimitation of these three types of measurements but not the elimination of any one.

It is fairly certain that the objective examination will displace much of the testing now done by the traditional written test or essay examination, particularly in the measurement of information. It is still an open question whether the objective test will prove adequate for the measurement of appreciational skills. For the time being we must choose between the subjective question which is most difficult of evaluation and the objective test item whose validity is not entirely assured.

There can be no ultimate conflict between the objective test and the standard test. As our skill in objective examining of the informal type increases, there is certain to arise a more critical attitude toward standard tests. This will have far-reaching results as there are many present signs of storm clouds hovering over the educator who accepts the standard test on faith. The standard test has thus far taken advantage of a "halo" arising from a largely mistaken belief that it represents a more "scientific" instrument than the classroom teacher can construct. This is not true in general. The best of existing standard tests do represent a degree of refinement not possible without extended experimentation. But the rank-and-file of such tests are readily equaled or bettered by the teacher who has mastered a little theory of measurement and who is seriously intent upon building valid and reliable examinations.

The author has drawn freely upon his two earlier books on the same subject, particularly the *Improvement of the Written Examination*. Grateful acknowledgment is made to

Scott, Foresman and Company for this privilege. There are a number of conclusions and recommendations in the present treatment which are in contradiction with previously published statements. The available experimental evidence in 1923-1924 when the *Improvement of the Written Examination* was being written was almost negligible. The past five years have brought forward a very respectable mass of empirical findings. If the author in his earlier writings guessed, and guessed wrong at times, it may be charged to the general lack of knowledge at the time.

The reader will undoubtedly be conscious of the fact that there is a considerable amount of repetition of certain ideas like the theory of sampling, need for reliability, measurement as ranking, etc. This was done deliberately under the feeling that such ideas are not familiar to most teachers and that repetition is one of the best means of emphasis. The author is quite aware that a more concise and logical, but less psychological, treatment might have been pursued.

Dr. Noel Keys, Associate Professor of Education, University of California, read all the manuscript and made scores of valuable suggestions. Dr. Hermann Remmers of Purdue University also read portions of the manuscript to its great improvement. The author's greatest indebtedness is to Dr. Ben D. Wood of Columbia University. Although geographically far apart, Dr. Wood and the author have been very close together in the directions which their studies have taken. At times they have quite independently arrived at the same conclusions. This has been a source of great satisfaction, particularly since Dr. Wood is recognized as perhaps the foremost investigator and advocate of the new-type examination. Less directly the author is indebted to his former teachers of educational and mental measurement, Doctors Lewis M. Terman and Truman L. Kelley of Stanford University. The World Book Company, the Macmillan Company, the Bureau of Public Personnel Administration,

and the State University of Iowa have permitted the quotation or reprinting of certain of their copyrighted materials. A great many public school teachers have contributed objective examinations used for purposes of illustration. Specific reference is given to these sources at the proper places. The *General Bibliography* at the end of the volume is the work of Mr. Sanford Siegrist, Mr. George Meyer, and the author. The numerous studies of the author's graduate students are given recognition throughout the text.

Miss Birdie Weisbrod and Mr. George Meyer read all the proof and checked many of the calculations.

It is the hope of the author that a great many classroom teachers and beginning students in education will take the trouble to go through at least Parts I and II of this volume and accept or reject the various points of view. It is not to be expected that the author is invariably correct in his statements. The important thing is to have thought through the issues presented.

G. M. R.

Berkeley, California
January 5, 1929

CONTENTS

PART I: THE ARGUMENT FOR OBJECTIVE EXAMINATIONS

CHAPTER PAGE

I. Points of View 3
Early ideas on examinations 3
The retention or elimination of examinations 8
The functions served by examinations 10
The principal kinds of examinations 18

II. The Criteria of a Good Test or Examination 27
Statement of the criteria 27
Validity 27
Reliability 40
Ease of administration and scoring 63
Norms or standards for evaluating test scores 66
Duplicate or equivalent test forms 66

III. Objections to the Traditional Examination 70
Investigations of teachers' marks 70
Investigations of regradings of the same papers 77
Studies on the reliability coefficients of examinations . . . 89
Reduction of subjectivity through scoring rules 101
Summary, discussions, and conclusions 106

IV. Advantages and Limitations of Objective Examinations . 112
Advantages of the objective examination 112
Limitations of the objective examination 120

V. Students' Attitudes Toward Examinations 130

VI. Relative Values of Standardized and Non-standardized Tests 138

PART II: HOW TO CONSTRUCT AN OBJECTIVE EXAMINATION

CHAPTER PAGE

VII. The Building of an Objective Test or Examination . . . 149
Drawing up a table of specifications . . . 150
Drafting the items in preliminary form . . . 153
Deciding upon the length of the test . . . 159
Editing and selecting the final items . . . 160
Rating the items for difficulty . . . 163
Breaking the items into equivalent forms . . . 164
Rearranging the items in order of difficulty . . . 166
Preparing the instructions for the test . . . 166
Making the answer keys or stencils . . . 172
Deciding upon rules for scoring . . . 184

VIII. Illustrative Types of Objective Tests . . . 188
Recall types . . . 191
True-false types . . . 194
Multiple-response (multiple-choice) types . . . 198
Matching exercises . . . 200
Analogies . . . 202
Rearrangement types . . . 202
Computations . . . 204
Constructions . . . 204
Identifications . . . 205
Reproductions . . . 206
Correction of errors . . . 206
Redundancies . . . 206
Map location . . . 208
Deduction of conclusions from premises . . . 209
Translations . . . 209
Miscellaneous and mixed types . . . 210

IX. Selected Complete Examinations . . . 213
Kern County examination for grade six . . . 214
Wyoming state examination in agriculture . . . 224
Rochester High School American history examination . . . 229
American Library Association examination . . . 241
Alameda High School literature test . . . 248
Test of musical accomplishment . . . 254

CHAPTER PAGE

X. Rules for Drafting Objective Test Items 265
True-false tests 265
Simple-recall tests 269
Completion tests 271
Multiple-response tests 274
Matching tests 276

PART III: EXPERIMENTAL AND THEORETICAL CONSIDERATIONS

XI. Experimental Studies on New-type Examinations . . . 281
Comparative validities 281
Comparative reliabilities 291
Comparative working times 306
Comparative difficulties 314

XII. Chance and Guessing in Recognition Tests 318
The mathematics of chance applied to tests 319
Experimental investigations 331

XIII. The Negative and Other Suggestion Effects in the True-false Tests 358

XIV. Examinations, Marks, and Marking Systems 369
The percentage grading plan 370
Grading by the normal curve 374
Measurement and ranking 392

PART IV: STATISTICAL TREATMENT AND INTERPRETATION OF OBJECTIVE TEST RESULTS

XV. Statistical Problems Related to Measurement 405
Summarizing a series of test scores 405
Determining the reliability of a test 412
Uses of the Spearman-Brown prophecy formula 420
The measurement of variability or dispersion 422
The accuracy of an individual score 428
What is a satisfactory degree of reliability? 433
The effect of heterogeneity or range of talent on reliability coefficients 441

PAGE

GENERAL BIBLIOGRAPHY 447
I. Books, monographs, and bulletins 447
II. Unreliability of teachers' marks, marking systems, etc. . . . 449
III. Comparative studies of old- and new-type tests 452
IV. Instructional uses of objective tests 454
V. Students' attitudes toward examinations 455
VI. Samples of objective tests 455
VII. Experimental and theoretical papers 459
VIII. Miscellaneous 464
IX. Selected textbooks on educational measurement 469
X. Selected references on statistical methods 471

INDEX 472

Part I

THE ARGUMENT FOR OBJECTIVE EXAMINATIONS

CHAPTER I

POINTS OF VIEW

EARLY IDEAS ON EXAMINATIONS

Historical and introductory. The past quarter of a century has witnessed the rise of educational measurement to the plane of conscious striving for objective, impartial, and comparative means for portraying the absolute and relative achievements of pupils. Prior to the beginning of the present century, teachers, although long familiar with examinations, did not view their tests and examinations as measurements in the present meaning of the word. Oral quizzing, Socratic or otherwise, had been from time immemorial a part of the daily classroom routine; in fact, at times, it was all of teaching. Formal written examinations are probably more recent than oral testing, but these date their origins many centuries ago; certainly formal written examinations were firmly intrenched in the educational system of China thirteen hundred years ago, and were familiar to Grecian and Roman teachers.

In America examinations appear to be as old as formal education itself. Horace Mann, as early as 1845, formulated a clean-cut concept of the written examination and its superiority over such older methods as the oral quiz.

The serious student of the art of examining will find a mine of delightful and valuable information about early examination methods in that most interesting volume by Professors Caldwell and Courtis entitled *Then and Now in Education; 1845-1923.*[1]

[1]Published in 1924 by the World Book Co., Yonkers-on-Hudson, N. Y.

Horace Mann and the written examination. Writing in 1845 Horace Mann argued that the new examination was superior to all other methods for the following reasons:

1. It is impartial.
2. It is just to the pupils.
3. It is more thorough than older forms of examination.
4. It prevents the "officious interference" of the teacher.
5. It "determines, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught."
6. It takes away "all possibility of favoritism."
7. It makes the information obtained available to all.
8. It enables all to appraise the ease or difficulty of the questions.[1]

These arguments were advanced in justification of what was really the first American school survey, that of the Grammar and Writing Schools of Boston in 1845. Horace Mann proceeds further to justify the "new" examination:

. . . it submits the same question not only to all the scholars who are to be examined, in the same school, but to all schools of the same class or grade. Scholars in the same school, therefore, can be equitably compared with each other; and all the different schools are subjected to a measurement by the same standard. Take the best school committee-man who ever exposed the nakedness of ignorance, or detected fraud, or exploded the bubbles of pretension, and let him examine a class orally, and he cannot approach exactness in judging of the relative merits of the pupils by any very close approximation. And the reason is apparent. He must propound different questions to different scholars; and it is impossible that these questions should be equal, in point of ease or difficulty. A poor scholar may be asked a very difficult one, and miss it. A good scholar may be asked a very difficult one, and miss it. In some cases a succeeding scholar may profit by the mistakes of a preceding one; so that, if there had been a different arrangement of their seats, the record would have borne a different result of plus and minus. The examiner may prepare himself as carefully as he pleases, and mark out the precise path he intends to pursue, and yet, in spite of himself, he may be thrown out of his path by unforeseen circumstances. But when the questions are the same, there is exactness of equality. Balances cannot weigh out the work more justly. So far as the examination is concerned, all the scholars are "born free and equal."

[1]From Caldwell and Courtis, *Then and Now in Education* (Yonkers-on-Hudson, New York: World Book Company, 1923), p. 37.

Suppose a race were to be run by twenty men in order to determine their comparative fleetness; but instead of bringing them upon the same course, where they could all stand abreast and start abreast, one of them should be selected to run one mile, and so on, until the whole had entered the lists; might it not, and would it not so happen that the one would have the luck of running up hill, and another down; that one would run over a good turnpike and another over a "corduroy"? Pupils required to answer dissimilar questions are like runners obliged to test their speed by running on dissimilar courses.

Again, it is clear that the larger the number of questions put to a scholar, the better is the opportunity to test his merits. If but a single question is put, the best scholar in the school may miss it, though he would answer the next twenty without a blunder; or the poorest scholar may succeed in answering one question, though certain to fail in twenty others. Each question is a partial test, and the greater the number of questions, therefore, the nearer does the test approach to completeness. It is very uncertain which face of a die will be turned up at the first throw; but if the dice are thrown all day, there will be a great equality in the number of faces turned up.[1]

Before commenting on the foregoing quotation, let us proceed further with the reproduction of these prophetic and discerning ideas of Horace Mann.

Suppose, under the form of oral examination, an hour is assigned to a class of thirty pupils; this gives two minutes apiece. But under the late mode of examination (the uniform written examination), we have the paradox that an hour for thirty is sixty minutes apiece. Now it often happens that a sterling scholar is modest, diffident, and easily disconcerted under new circumstances. Such a pupil requires time to collect his faculties. Give him this, and he will not disappoint his best friends. Debar him from this, and a forth-putting, self-esteeming competitor may surpass him. In an exercise of two minutes, therefore, the best scholar may fail, because he loses his only opportunity while he is summoning his energies to improve it; but give him an hour, and he will have time to rally and do himself justice. It is one of the principal recommendations of this method, indeed, that it excludes surprise as one of the causes of failure, and takes away the simulation of it as an excuse.

And again:

It sometimes happens that when an examiner has brought a class or a pupil to a test-question—to a point that will reveal their condition as to

[1] *Loc. cit.*, pp. 39-40.

ignorance or knowledge—the teacher bolts out with some suggestion or leading question that defeats the whole purpose at a breath.[1]

Comment on the views of Horace Mann. It must be disconcerting to the modern writer on educational measurement to read these paragraphs from the pen of Horace Mann nearly a century ago. This volume by Caldwell and Courtis carries an eloquent reprimand to those of us who, in writing on topics related to tests and testing, have, through ignorance of earlier thought, imagined that the science of educational measurement is wholly novel and quite recent. Dr. J. M. Rice has been rather generally credited with being the father of educational measurement, but these quotations show clearly that certain of the essential ideas of the measurement of classroom products antedated these commonly accepted pioneers—Rice, Thorndike, Courtis, Stone, *et al.*

The author hastens to warn the reader that these appreciations of the insight of Horace Mann must not be carried too far on the wave of enthusiasm. After all, Horace Mann's examination—his "idea in mind"—was not the standard test of today nor even the modern teacher's concept of an adequate objective and impartial instrument of evaluation. The author is no student of the history of education; research will in all probability reveal even more remote writers who were the source of stimulation to the thinking of Horace Mann. Present comments are not directed at the establishment of the ultimate sources of our ideas on measurement, but rather, at the expression of a proper humility and an acknowledgement that the fundamental thinking on the merits and limitations of written examinations is an older story than current textbooks on tests and measurements have given us to believe.

Analysis of the ideas of Horace Mann. The quotations from these writings of Horace Mann will repay the reading

[1] *Loc. cit.*, pp. 40-41.

again. There are certain ideas imbedded there which form the basis of current practices in educational measurement. If the reader will follow the suggestion of re-reading these quotations analytically, he will find the following fundamental propositions boldly stated:

1. Examinations should be written rather than oral. (This of course refers to those final evaluations, which, in the custom of 1845, were delegated to the school committees.)
2. The questions presented should be uniform for all.
3. Uniform written examinations are more economical of time than are oral, individual examinations.
4. Uniform written examinations are *longer* in effect; i. e., they "sample more widely" (to phrase the statement in the modern terminology).
5. Examination questions differ greatly in difficulty, and such differences operate to obscure the real differences in ability among pupils.
6. Oral examinations tend to be unsystematic, and to be deflected from the aim of the examiner by unforeseen circumstances.
7. Uniform examinations place all students under the same conditions; i. e., they approximate an experimental situation, rather than the analogy of Horace Mann where twenty men ran a mile race over unequal courses.
8. Any examination is a limited sampling of a pupil's knowledge and skill; the larger the number of questions the fairer the test.
9. There is a marked chance element in success or failure on an examination. (The use by Horace Mann of the analogy of the throwing of dice is a striking forerunner of much recent controversial literature on such tests as the true-false and multiple-choice.) Mann's statements of the outcomes of repeated throws of a die and the long-time stabilization of results cannot fail to interest the student of modern literature on examination methods.

10. Oral examinations are relatively less reliable because of the greater tendency to emotional disturbances as compared with written tests.

11. Examinations should be freed from the inadvertent and very human tendency of teachers to assist pupils with the answering of the questions.

12. Examinations should be of considerable length. (This is implied rather than stated when Mann chose sixty minutes as the length of an examination in contrasting the scope of oral and written examination calling for one hour of time.)

The writer is somewhat disturbed by the danger of reading more into Mann's statements than the latter implied. Of this the reader must be the judge; in most instances Horace Mann seems to state the points without ambiguity, although present knowledge has refined and greatly extended these principles.

Regardless of the degree of parallelism of Horace Mann's ideas and the recent development of standardized and unstandardized objective (or "new-type") examinations, these quotations, comments, and analyses form an excellent starting-point for the series of discussions which have been grouped in this chapter under the heading, "Points of View."

THE RETENTION OR ELIMINATION OF EXAMINATIONS

The justification for examinations. Since there are so many vociferous critics of examinations, (by no means confined to public-school pupils and rebellious college undergraduates) it is fair to raise the question whether such measures are to be justified at all. Certain it is that the arguments pro and con rest largely upon opinion. There is no convincing evidence that examinations are essential to the complete procedure of instruction. There is similarly no indisputable evidence that they are not. For this reason

alone, the present treatment of examinations is justified in avoiding this highly controversial issue.

In the second place, no one even reasonably well informed about recent developments in education can doubt that the written examination is taking a new lease on life. A quarter of a century and less has witnessed the rise of the standardized examination, and still more recently the unstandardized, informal, teacher-built, objective test.[1]

The whole modern emphasis on the psychology of individual differences and the attendant problems of measurement is too recent and too firmly intrenched to make probable any lessening of the esteem in which examinations are held. Aside from these newer developments the examination system is firmly established, and although it will certainly remain under fire from those who look on it with disfavor, present indications point strongly to the fact that the main effort will be expended in perfecting it along directions suggested by recent experimentation.

In general, it will be more fruitful to attempt to better a practice which consensus of opinion still favors before considering the more drastic step of complete elimination. The burden of reform always falls upon the reformer. Whether the written examination holds its place in the curriculum from merit or from precedent, those who would do away with it must assume the obligation of experimental demonstration of the futility of examinations. The critics of examinations

[1]The type of examination with which this book deals principally is variously termed "informal," "new-type," "unstandardized," "short-answer," and "objective." None of these names is particularly fortunate since each emphasizes but a single one of several important attributes of such examinations.

On the whole, the author prefers the term *objective examination*, believing that it emphasizes the most important single point of difference between two contrasting types of examination, i. e., objectivity or freedom from personal opinion in evaluating examination results.

The term *new-type examination* has been very largely employed by Dr. Ben D. Wood, a foremost student of examination methods. This designation unfortunately seems certain to lose its meaning with the passing of time. (It should be recalled that Horace Mann used an almost identical terminology in 1845, and already we are proposing a "newer" type of test.)

The present volume will use several of these designations interchangeably, following the practice of most writers.

have mostly argued a priori; they have marshaled little or no experimental evidence, and the present temper of professional education draws it constantly further from conviction by argumentation and constantly nearer to conviction by experimentation. The worth of examinations is open to crucial experimental determination, and nothing short of this will be convincing in the long run.

For reasons advanced above, in part, it is justifiable to avoid completely the broader question of the retention or elimination of examinations, and to study instead the possibilities of improving our devices for measurement of school accomplishment.

THE FUNCTIONS SERVED BY EXAMINATIONS

Classification of the purposes of examinations. Although a great many specific functions have been claimed for the written examination by one writer or another, these may for present purposes be grouped as four, as follows:

1. Motivation of the learning of pupils.
2. Maintenance of standards of accomplishment.
3. Training in the use of the English language.
4. Measurement of accomplishment.

Examinations for motivation. It is unfortunate that we have so little direct information as to the motivating effect of examinations. That examinations do have this value has been tacitly agreed but never proved. In spite of this dearth of proved fact, it does seem reasonable to suppose that pupils strive for somewhat greater and somewhat more permanent mastery when they realize that searching examinations may be expected at a later date. If this conclusion is true, certain reforms in the examination system might greatly increase the value of the examination as a motivator. We might argue somewhat as follows:

1. The motivating value of an examination varies with the esteem in which it is held by pupils. The more impartial and objective the examination marks, the more meaning they will have for the pupil.

2. Examinations should come at frequent intervals, and should not be confined to end-of-semester and end-of-year testing. To examine extensively but infrequently delays the day of reckoning so long as to make the goal too remote to stimulate the pupil. It also encourages the cramming attitude, which is of doubtful value.

3. Where tests and examinations punctuate the teaching at frequent intervals, it is possible for the pupil to keep cumulative, graphic records of his achievement. Such records form one of the strongest forms of motivation which we know today. Experimental psychology has repeatedly demonstrated that output with knowledge of results is markedly greater than is the case where the learner is kept in ignorance of his successes and failures.

4. When tests are of a detailed, specific, and diagnostic character, pupils cease to regard them as drudgery but come to depend upon them for guidance in remedying their weaknesses and as preparation for future opportunities to better past records.

The traditional examination provided a not altogether wholesome attitude on the part of pupils. As goals they were too remote. They called for an excessive expenditure of physical energy in writing. They represented an undue balance in the apportionment of time; too much for writing, too little for thinking. There was a considerable and growing distrust of their accuracy in differentiating among the abilities of pupils.

If examinations in the past have failed to serve as motivators, the fault is possibly due to the kind of examination. If the arguments advanced here are sound, it appears hope-

ful to attempt to increase the motivation value of the examination through reforms in the examination system itself.

Maintenance of standards of work. Many school supervisors feel that examinations and tests set by them offer a good means of control of standards of work by different teachers. This belief has led to the practice of conducting uniform city-, county-, and state-wide examinations. Such practices appear to be losing ground slowly, although almost fifty per cent of the individual states do have uniform state examinations in at least the eighth grade. In some states the county boards of education are the examining bodies. It is in the cities that uniform examinations have lost most ground, although some cities are returning to something like uniform examinations, except that modern objective tests are employed instead of the older essay examinations. Standard educational and mental testing has provided a more secure basis for rendering instruction sufficiently uniform from one classroom to the next.

This brings us to a related problem: that of evaluating the efficiency of the teacher by tests of her pupils' accomplishments. It was on this point that Dr. J. M. Rice drew so much fire from the National Education Association at the beginning of this century. It was first thought that the standard test would serve to evaluate teaching upon the principle that if pupils showed high accomplishment, the teaching was good, but if pupil achievement was low, teaching was, perforce, unsatisfactory.

The dangers inherent in this point of view were soon exposed. The standard test method made no allowances for differences in pupils' mental equipment, the most important single factor controlling the rate of learning yet found. It was soon realized that standard tests were unadaptable to local conditions, that they were open to abuses through coaching, that they were not as unerring guides as first

supposed, and that they often misfired in reaching the essential activities of the classroom.

There is a legitimate place in the scheme of supervision for uniform examinations. If the general merit of a series of uniform examinations could be established, the periodic application of such tests would accomplish much by way of equalizing instruction and defining objectives throughout school systems. Where objectives and aims can be translated into concrete test situations, supervision through locally constructed tests is far more economical than personal supervision. Such tests must obviously have a high degree of accuracy; they must confine themselves to essential outcomes without placing limitations upon the means of arriving at these outcomes; they must parallel the curricular units with exactness; they must be numerous enough to cover all major curricular units; and they must be constructed in duplicate and equally difficult forms so that they do not need to be repeated sufficiently often to lay them open to abuse through coaching and cramming.

Further discussion of these points will be postponed for Chapters II and VII.

Examinations as training in language. One of the strongest arguments for the traditional examination has been its reputed value in teaching pupils to organize their ideas and to place them on paper in good English. If written examinations in the past have served well such an obviously important function, a change to the mechanical objective examinations will be a real loss to linguistic training.

But what are the probable facts? Do the conventional examinations contribute importantly to the teaching of the English language?

The conviction is slowly gaining ground that the value of written examinations in establishing good language habits is largely illusory. Since this book advocates replacement of

much of our examination system by the new-type objective tests (of little or no value in language training), its recommendations are certain to be opposed by those who think the examination supplies useful language drill. In such a case the best defense is to attack. The arguments against the written examination, as contributing significantly to good English habits, may be marshaled somewhat as follows:

1. Most teachers have noted that final examinations show a quality of diction, grammar, and spelling markedly inferior to the products of the regular English, composition, and spelling classes. Some teachers are convinced that the written examination has a negative or destructive value in English training.
2. The actual conditions of the examination period are unfavorable to good linguistic expression. The pupil is usually required to write five or ten long questions, very often taxing his speed to finish at all. He has little or no time to reflect upon his literary style. He will be lucky to get down the facts in the allotted time.
3. The pupil realizes, consciously or unconsciously, that the paper will be graded upon facts, not style. He senses that his teacher wants to find out what he knows about geography, physics, etc., and he acts accordingly.
4. Language habits are complex. They, like all other habits, are built up slowly and consciously. They do not arise by magic. It is unlikely that they will arise as by-products of frenzied efforts at setting down facts in limited time.

The reader must judge of the soundness of such arguments. If true, the ordinary written examination contributes nothing to language development, and if, further, the new-type examination comes into general use, it will be necessary to seek some method of providing for the measurement of the power to organize and express ideas on paper. Possible solutions of this problem are:

1. Give up all attempts to make the examination period serve the purposes of language training and make provision for such needs elsewhere.

2. Divide all (or certain) examinations in two parts, somewhat as follows:

Part I: Objective (completion, true-false, etc.) in character; suggested number of items, 75 (as a minimum); time allowed, 20-30 minutes; total credit allowed, 75 points out of a total of 100.

Part II: Discussional (essay-type) in character; suggested number of questions, 1 (or at most 2); time allowed, 30 minutes (as a minimum); total credit allowed, 25 points out of a total of 100.

The division of credit as 75:25 is based upon the logic that a composite score on the two parts should not be rendered too inaccurate by allowing too much weight to the second part of the examination (which is not open to accurate marking). It should also be noted that 30 minutes per question is suggested as the time allowed under Part II. The reason for this will soon be made apparent.

Part II should be accompanied by instructions to the pupil substantially as follows:

Instructions: 1. In this part of the examination you are not to be graded upon the number or accuracy of the facts which you put down.

2. Your mark will be based upon the following factors: (*a*) evidence of thought, (*b*) good sentence and paragraph structure, (*c*) grammatical and dictional errors, (*d*) spelling, etc. In other words, consider that you are writing a theme or composition for an English class.

3. Read the question (or questions) at least twice before you start to work.

4. Do not attempt to write anything until you have made a *written outline* of what you are going to say.

5. As you make your outline, think through what you intend to write as your answer to the question.

6. *Do not hurry!* Spend at least half of the time allowed in thinking and planning your outline.

7. *Remember:* You are not to be graded upon the facts about geography (or subject in question). Your mark will be based upon your ability to express your thoughts in correct English.

Now it follows that a teacher electing to conduct such an examination must "play the game fairly" with her pupils. The pupil must not feel any pressure of time. The teacher must steel herself to avoid, as far as possible, marking the replies upon a basis of subject-matter, and she must take her time in evaluating such a paper. The marking will be highly subjective at best.

The author, although perfectly serious in the foregoing proposal, does not believe that many teachers will adopt the type of examination suggested for Part II. The honest-minded ones will see that our proposal is consistent with an unprejudiced effort to make the written examination function in language training. The majority of us will be more likely to conclude that "If I have to go to all this bother, I'll find some other way of teaching English. After all, examinations are intended to measure pupils' abilities, not to teach them the English language." We may not condone such an attitude —but is it not the expectancy?

Examinations as measurements. The measurement of achievement has been admittedly the principal reason for examinations. This idea is undoubtedly sound. It may require certain qualification, but all seem to be agreed that the first purpose of a test or examination is that of ascertaining the degree to which individual pupils have profited by instruction.

Although we have referred to *measurement* as a single idea or term, it really includes a number of fairly discrete purposes, viz.:

1. Measurement of general or all-round ability in a school subject. These are usually comprehensive final (semester or year) examinations designed to show the pupil's general

grasp of the subject and to stimulate him to review. Such tests or examinations should represent the "high points" and should help to leave the pupil with a bird's-eye view of the subject.

2. Diagnosis of specific strengths and weaknesses in instruction and the profit from instruction. Such tests are usually given at the time of completion of each important teaching unit. They are detailed, and they must be very reliable if the diagnosis is to have meaning for individual pupils. Results from such diagnostic tests should lead to two general outcomes:

(*a*) Improvement of the instruction of the teacher.

(*b*) Guides to remedial or corrective work for the pupil.

3. Examinations for prognosis, placement, sectioning, etc. Although standard educational and mental tests are more often used for this purpose, informal objective tests have great possibilities here. There are dozens of questions to be answered in handling pupils in a modern school, e.g., admission to high school or college, placement in proper grades of transfers, sectioning into ability groups, educational and vocational guidance, prediction of future success, etc.[1]

The question of when a pupil has been measured accurately is a principal theme of this volume. Several chapters further on we shall be in a better position to discuss examinations in the light of the theory and practice of measurement. There it will be shown that measurement is never complete, but merely samplings of ability. "Old" and "new" types of tests will be studied in the light of the criteria which good examinations must meet. There is a new vocabulary, that of the professional student of examinations, which must be assimilated and absorbed into our thinking before we can hope to approach the ins and outs of various types of mea-

[1]*See* P. M. Symonds, *Measurement in Secondary Education* (New York: The Macmillan Company, 1927), pp. 1-2.
G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction* (Yonkers-on-Hudson, New York: World Book Company, 1927), pp. 8-44.

surements in a critical and analytical fashion. For these reasons further discussion of this point is unwarranted at this time.

THE PRINCIPAL KINDS OF EXAMINATIONS

A classification of examination methods. Four types of measurements exist side by side in the modern school. These are:

1. Oral questioning
2. The traditional examination[1]
3. The standard test
4. The objective or new-type examination

There are other means of evaluating school results, but the four types mentioned are the most important.

With such a variety of methods open to the educator, and with so little of a final character known about their relative merits, the only course open to us is to consider both the logic and the growing body of experimental findings supporting or undermining the value of each. This is the task of this volume in its entirety, but a few brief comments will help to give a point of view.

The oral examination. Strictly speaking oral questioning does not usually constitute an examination. Oral examinations are sometimes employed, but, with Horace Mann, we doubt their value for the more serious and final determinations of achievement. This is no argument against oral

[1]The rise of objective or new-type examinations makes necessary a distinction between the long-established form of test and the more recent and more objective type of examination. The former has come to be known as the *traditional examination* or the *essay examination*. The traditional examination needs no definition. It is the examination which we all recognize as consisting of five, ten, or more questions, beginning most often with "State in full," "Describe," "Tell what you know," etc. The pupil is free to write what he chooses as a response to the stimulus question. It is to be contrasted in its mechanics with the newer objective examination in that the latter calls for underlining, crossing-out, checking, etc., instead of discussion. The traditional examination cannot be scored mechanically by keys or stencils but must be evaluated subjectively by competent persons.

questioning. In many ways the teacher's daily questioning of her pupils is of far more fundamental importance than her final written examination. The point is that oral questioning is more logically a part of initial instruction than of final measurement, assuming that there are at least five roughly distinguishable phases to the complete act of instruction, as follows:

1. Initial presentation of materials to be mastered. This phase consists of setting problems to be solved, textbook readings and discussions, teachers' comments on persistent difficulties in learning, etc.
2. Drill to support and fix the temporary mastery gained under the first phase of instruction. This may be drill proper or it may mean applications and reviews.
3. Diagnostic measurement at the period when phases one and two are thought to be complete.
4. Re-teaching or remedial instruction upon any weaknesses revealed under the third phase.
5. Final measurement and evaluation of a more general and less detailed character than that of phase three. This constitutes the final survey of achievement and leads to a judgment as to whether the individual or class is ready to proceed to new work.

It should be noted that certain of these phases are less prominent than others at times, the relative emphasis varying with the character of pupil, teacher, textbook, subject, motivation, etc.

Oral questioning plays its greatest role in the first, second, and fourth phases of instruction as presented above. It is primarily instructional; its value for measurement is more subordinate. Oral questioning as an art has a long history and a considerable literature. It is worthy of more experimental study than it has received to date.

The traditional written examination. The familiar discussion- or essay-type examination has long been the principal reliance of the teacher in evaluating pupil accomplishment. It is probably the most frequently employed examination at present, although this dominance shows signs of breaking. Its advantages cannot be stated more clearly than our previous quotations penned by Horace Mann almost a century ago. Its weaknesses are numerous, but this is not the place for the discussion of such limitations.

It is sufficient to point out that the essay examination suffers from one major defect not inherent in the standard test or the newer objective examination; viz., *experience and experiment have shown that the results of an essay examination cannot be evaluated fairly by human minds.* Its inaccuracies are those of the human mind and the human prejudice. Such examinations seemingly cannot be freed from the personal equation.

Our chief interest in the present chapter is in broader points of view about examinations. The mark on the conventional examination cannot fail therefore to interest us as students of examinations. To the degree that an examination mark or grade reflects the knowledge, attitudes, and prejudices of the marker of that examination paper, the examination is not a true measurement since all are surely agreed that it is the accomplishment of the pupil which is to be measured. If, as we shall see later, the same pupil's paper is graded all the way from 40 to 90 (as many investigators have found), there is but one conclusion to be drawn; viz., *the pupil has not been measured.* To be at the same time a "40" pupil (a dunce) and a "90" pupil (a candidate for the class valedictorian) is not only unthinkable but palpably untrue! Such a finding raises the suspicion that he is *neither*, a conclusion that can well be supported on the ordinary logic underlying our basic theorems of probability.

To be taken at face value, any examination result must meet many stringent criteria, and one of these is that it be a measure of the *pupil*—not the teacher, not his class, and not the school system. Yet it must be admitted by any fair-minded student of the literature that the traditional examination is prone to tell us as much, or almost as much, about whom the pupil had for a teacher as it does about the educational equipment of the pupil himself. We shall not trouble to prove this assertion at this time.

The technical term for this weakness in the common essay-type examination is, in our modern educational terminology, *subjectivity of marking*. It was as a relief from this admitted weakness that the standard test and the objective examination were introduced. How adequate the remedy will prove to be cannot be foretold here, although we may study the evidence, combine this evidence with our logical and experimental deductions, and finally arrive at a tentative point of view. This will have to be the task of succeeding chapters.

The standard test. Standardized examinations have just completed the first quarter century of their existence. From a few pioneer attempts by Rice, Thorndike, Stone, Courtis, and others in the fields of spelling, arithmetic, and reading, the movement has grown until conservative estimates place the total number of available tests and scales at at least five hundred; there are probably considerably more. It is impossible to secure even approximate estimates of the numbers of standard tests administered annually. There are several educational tests whose sales have passed the million mark annually. In one or two cases, two million is a more nearly correct figure. The total number of standard tests sold during the past year is probably at least twenty million, possibly somewhat more.

These figures, estimates as they are, point to the importance of standard tests as measures of the results of teaching. It seems certain that the curve of the use of standard tests is rising more rapidly than is that of the increase in school population.

The standard test was introduced to serve several purposes. These are not in all respects a prime concern of the present treatment, but they serve to orient our thinking about tests and examinations in general. The principal aims of the standard test may be listed as follows:

1. They (as the name implies) represent an attempt to control or standardize the conditions of the examination period with respect to directions, time allowances, method of responding, etc.
2. They are objective or impartial; i. e., the personal equation of the examiner is minimized or eliminated—minimized in the administration, and eliminated almost or quite completely in the scoring of the examination.
3. They provide norms or standards (as the name further implies) by which the scores of individual pupils may be evaluated and interpreted in the light of facts. Such facts are the performances of large numbers of supposedly typical pupils on the same tasks.

These aims can all be attained to degrees commensurate with the practical needs of education, the third aim being the most difficult, and, on the whole, decidedly the least important.

Against these advantages of the standard test may be set certain more or less fundamental limitations which are briefly enumerated below.

1. Standard tests are inflexible and cannot be closely adapted to the idiosyncrasies of local school conditions.

They are of necessity general enough to meet moderately well a wide variety of curricula.

2. In view of the foregoing, they need constant supplementation in a complete measurement program.

3. Standard tests are somewhat expensive. The range of prices varies from about one cent per pupil to at least ten cents per pupil. This, of course, is a practical limitation, not a theoretical one. It should also be noted that there is considerable correlation between cost and worth. As is the case with all commercial products, tests are sold in a competitive market, and costs are reckoned upon the basis of the expenses of production.

4. The majority of standard tests are of little value. A large number are nothing more than "examinations with norms," produced by persons without special training or knowledge of test construction. If one hundred of the best were selected and the rest destroyed, the loss would be negligible.

Only the first-mentioned of these limitations of the standard test is serious. The others may be overcome by careful selection, by the planning of measurement programs, and by efficient school budgeting. It would appear to be impossible to adopt the standard test as the sole element in a measurement program. It might well repay the cost, but it is to be doubted whether local needs could ever be met satisfactorily. Both the traditional and the new-type examination are free from this limitation of non-adaptability to local school curricula.

The objective test or examination. For the present we will define objective tests by means of a few test items illustrating several of the principal types.

1. (*Simple recall*) The Senate and the House of Representatives together form the United States .....

2. (*Completion*) Oxygen is often prepared by heating potassium chlorate together with ..... which acts as an accelerator of the reaction. Such substances are called ..... The formula for potassium chlorate is ..... Water may be decomposed by the electric current to oxygen and ..... in the ratio, by volume, of ..... to .....

3. (*True-false*) The rainfall is generally heavier on the eastern than on the western slope of north-and-south ranges in the path of the westerly winds. TRUE FALSE

4. (*Multiple-choice*) A reduction in price for buying in large quantities is called a **commission discount dividend mortgage revenue**

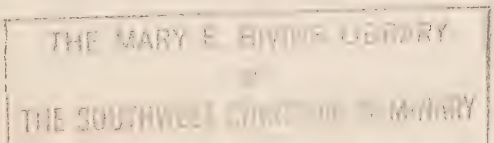
The objective or new-type test is essentially a hybrid. It represents the objectivity of the standard test without the refinements of experimental study and standardization. Because careful standardization is not attempted, it may be produced almost as cheaply as the traditional examination, although at a much greater expenditure of time and energy. It (like the traditional examination) has a high degree of adaptability to local conditions. The lack of norms is a limitation, but no more serious a one than is the case with the older forms of examinations. The relatively smaller degree of refinement, as compared with well-made standard tests, can be compensated for very largely by increased length. This point will be discussed at length in later chapters.

No data are available as to the extent to which objective, teacher-made tests are being utilized in our schools. There can be no doubt that these types of measurement are increasing in favor even more rapidly than are standard tests. Hailed a half-dozen years ago as "a new type of examination," they are regular routine in thousands of progressive schools. They show signs of their inroads into the practices of even the most conservative examination bodies. The New York Regents are conducting extensive experiments with these new tests. At least two of the states giving uni-

form state examinations are using objective tests wholly or in part.[1] There are doubtless others which have not come to the attention of the author. Certain school systems like Atlanta, Denver, Detroit, Los Angeles, Rochester, St. Louis, and many others are developing extensive batteries of such tests. In some states where uniform examinations are set by county boards of education, certain counties are now developing series of objective tests for the elementary school subjects.[2] No special attempt has been made to collect adequate statistics on the use of objective tests throughout the United States. The specific references are those which come to the author's mind at the moment of writing. In a recent competition for the construction of objective tests, about 400 examinations were submitted for consideration for the prizes offered.[3] Chapter VIII presents certain summaries of the findings from this competition.

Looking back over the five-year period in which the author has been engaged in studying experimentally one phase or another of examination construction, the story of the progress of objective examination methods is ample grounds for predicting that something like the objective examination, when perfected, will be the principal reliance of the classroom teacher for the next few decades to come. There will doubtless be a place for the traditional examination in the future, but it seems likely that it will tend to become a last resort, to be employed when other methods are not at hand. It may be possible to perfect the ordinary examination so as to control some of its vagaries, but progress to date leaves small reason to hope for marked success.[4] As has

[1]New Jersey and Wyoming.
[2]E. g., Lewis County, New York and Kern County, California.
[3]The results of this competition are published in G. M. Ruch and G. A. Rice, *Specimen Objective Examinations* (Chicago: Scott, Foresman and Company, 1929). This volume presents thirty-five of the best examinations in a total of nearly four hundred submitted. The prize-winning examinations will be found to be both interesting and valuable to the teacher just beginning to employ objective examination methods.
[4]*See* Chapter III, pp. 101-106.


THE MARY E. RIVER LIBRARY
THE SOUTHWEST

been said, the old- and the new-type must be regarded as complementary and not antagonistic; the latter type of examination will exclude the former more or less completely for informational subjects, and the former will doubtless continue to hold a place in the measurement of expressional and appreciational subjects.

CHAPTER II

THE CRITERIA OF A GOOD TEST OR EXAMINATION

STATEMENT OF THE CRITERIA

The "ear marks" of a good examination. The criteria for use in the construction of standard and unstandardized tests naturally differ somewhat. The following have been selected as sufficient for the main outline in understanding the theory and construction of school examinations of the objective type. We may therefore proceed to outline the principal criteria of a good test or examination as follows:

I. Validity
II. Reliability
  A) Objectivity
  B) Extensity or adequacy of sampling
III. Ease of administration and scoring
IV. Norms or standards for evaluation of results
V. Availability of equivalent or duplicate forms

VALIDITY

Definition of validity. The most important single fact which can be known about a test or examination is the degree of *validity* which it possesses. Validity may be defined variously; these separate definitions constituting, collectively, the ideas incorporated in the term.

1. Validity is the degree to which a test or examination measures what it is intended to measure.

2. Validity is the general worthwhileness of an examination.

3. Validity refers to the care taken to incorporate in a test or examination those elements or items which are of prime importance, and to the pains taken to eliminate the non-essential.

4. Validity is in general the degree to which a test parallels the curriculum and good teaching practice.

5. Validity refers to the value of the test for measuring specific abilities in an accurate fashion, and a test ceases to have validity when applied to the measurement of abilities for which it was not intended.

To these we must add, for the sake of avoiding misunderstanding, and not exactly by way of definition, the fact that validity is a broader term than reliability (defined later), and that validity includes reliability. That is, *a valid test must of necessity be a reliable test.*

The nearest synonyms for validity are "goodness," "general merit," and "worthwhileness."

It will add still further to our concept of validity to review the means by which validity is guaranteed in a test. Some of these means must perforce be reserved for Part II of this volume, "How to Construct an Objective Examination." The methods of validating standard tests are very instructive, and, although not entirely applicable to the purposes of objective test construction, they will help to define our terms at this time.

Principal methods of validating tests. These may be stated as follows:[1]

[1]The following outline is taken with changes and additions from G. M. Ruch, *The Improvement of the Written Examination* (Chicago: Scott, Foresman and Co., 1924), pp. 14-16.

For fuller and more technical accounts (with special reference to standard tests), *see:* G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction* (Yonkers-on-Hudson, New York: World Book Company, 1927), pp. 301-328, or W. S. Monroe, *The Theory of Educational Measurements* (Boston: Houghton Mifflin Co., 1923), pp. 56-105.

1. By judgments of competent persons.
2. By analysis of courses of study or textbooks.
3. By harmonizing with the recommendations of national educational committees or other recognized bodies on curricula, courses of study, minimum essentials, etc.
4. By experimental studies of social utility (such as the Horn and Thorndike studies of the most frequently used words, the Ashbaugh and Horn studies of spelling lists, the studies of Wilson, Woody, *et al,* on the arithmetic needs of business, etc.)
5. By studies of the most frequently recurring errors.
6. By the computation of the percentages of pupils answering each item correctly at each successive age or grade level.
7. By correlation against an outside criterion.
8. By combinations of the above methods.

Inspection of the foregoing list will show that the first three reduce to the single criterion of expert opinion. Textbooks, courses of study, and national committee reports are not usually to be regarded as resting upon experimental bases. Criteria four to seven are experimental in character, but these methods tend toward greater refinement than is usually possible or necessary in building informal classroom tests of the objective type.

Methods 3, 4, and 5 influence examination construction indirectly; i. e., curricula should embody all such empirical data, and the content of tests and examinations should follow closely the curricula. The *Yearbooks* of the National Society for the Study of Education and the Department of Superintendence of the National Education Asssociation have placed these and related studies at the disposal of teachers.

Methods 6 and 7 belong in the field of standard test construction. Such techniques cannot be applied directly in building informal tests. The teacher who wishes real insight into such factors as validity and reliability will be repaid for

some study of validation methods as applied in the construction of our best mental and educational tests.[1]

The teacher's examinations will ordinarily be validated by a combination of three criteria: (1) the local course of study, (2) analysis of the textbooks employed, and (3) her judgment on points of emphasis, inclusion, and exclusion. The guiding principle in validation should be: *the tests must parallel the actual teaching.* We may emphasize this point still further, and somewhat differently, by saying: *any test must represent an extensive sampling of the materials of instruction.*

Suggestions for validating tests. Part II of this volume will set up a practicable procedure for the actual construction of an objective examination. In particular, a plan for making a "Table of Specifications" will be presented. It is not the purpose to enter upon such a discussion here, although brief mention of the value of such a table or outline will help to clarify further our concept of validation. The following suggestions will aid the conscientious teacher in constructing good (valid) tests.

1. In the course of regular teaching, make a practice of jotting down good test items (questions) as they occur to you. This will save "racking your brains" when you come to examination time.

2. Place these test items on small bits of paper; 3x5 library cards are best. Make a file of these questions. Secure a filing case and keep these cards. This filing system allows the insertion of new items, the discarding of unsuitable ones, and easy sorting and arranging of test materials. When test items are written or typed consecutively on paper, revisions and alterations often necessitate laborious recopying. A card may be inserted or thrown away without entailing any other alterations.

---

[1]See Ruch and Stoddard, *op. cit.*, Part IV, pp. 301-375, or Monroe, *op. cit.*, pp. 56-105.

3. When the time comes to build an examination, draw up a Table of Specifications, as directed in Chapter VII. This will tend to guarantee a defensible balance of emphasis, freedom from non-essentials, and the inclusion of all important topics.

4. After the test is given, ask the pupils to suggest items that were ambiguous, misleading, or not understood. You will find that from 5 to 10 per cent or more of the items were not well worded. These must be revised or thrown away.

5. Where possible, try to have one or two other teachers criticize your test items and rate them for difficulty.

6. The validity of a test is raised by having the items of a proper degree of difficulty. Items passed by every child or failed by all contribute nothing to the test. The average of two or three teachers' judgments of difficulty is better than the judgment of one, and often is a close approximation to the truth.

7. The validity of a test is increased by having the easiest items first and the hardest ones last.

These suggestions can be carried out in practice without great expenditure of time or energy. They will go far toward guaranteeing a high degree of validity to a test or examination. There are more exact methods of validation which may be employed if still more accurate and worth-while tests are desired. These are described in the following section.

The validation of individual test items. In validating standard tests, validity and reliability of individual test items are experimentally determined by giving the preliminary tests to hundreds of pupils in different school grades. The *percentage of pupils passing each test item* is then computed. If the percentage of successes on each item rises sharply and uniformly from one grade to the next, the item is held

to be valid and reliable because it discriminates between different levels of ability. If the percentages of successes rise and then fall (and perhaps rise again), the item is thrown away because it is erratic in its behavior.

The following principles govern the final selection (validation) of test items. Certain of these principles apply *strictly speaking* only to standard test construction, although the teacher interested in making informal objective tests may occasionally wish to employ these to better her tests. In the main, the value of these principles will be in the nature of the further definition of validity and test validation.

1. Faulty wording (poor sentence structure, ambiguity of meaning, etc.) lessens the validity of a test item.
2. Test items which can be answered by none of the pupils (100 per cent failures) are valueless, and hence invalid.
3. Test items which are answered by all pupils (100 per cent successes) are valueless, and hence invalid.
4. "Tricky" or "catchy" questions are usually unsatisfactory.
5. A few very easy questions, even if passed by all, are justified at the beginning of the test for the sake of encouragement and motivation of the pupils, although such questions may not measure, i. e., discriminate differences in ability.
6. Full and simple directions, the generous use of samples, and fore-exercises (preliminary practice questions, not counted in the score) increase the validity of the test items.
7. Arrangement of the test items from easy to difficult increases both the validity and reliability of the test. The easiest item should come first and the most difficult last, with all intervening items arranged in order of increasing difficulty.

The first of these principles needs no comment except to emphasize that teachers must be on their guard to prevent loose and ambiguous statements of test items from ruining

otherwise valid materials. Weidemann[1] found violations of forty-six rules of grammar, punctuation, diction, spelling, etc., in a study of a number of true-false examinations.

The second and third principles together are intended to guard against the inclusion of worthless materials upon a basis of too great ease or difficulty. Consider three tests as follows:

Test A: 100 items, each answered correctly by every pupil.

Test B: 100 items, each failed by every pupil.

Test C: 100 items graduated in difficulty from items passed by all to items failed by all.

Every pupil would receive 100 on Test A. Every pupil would receive 0 on Test B. On Test C on the other hand, the scores of a class would vary from something above 0 to almost 100. Good pupils would make high scores and poor pupils low scores. We would say that Test C *measures* achievement but that Tests A and B are worthless (invalid) as *measures*.

It might be answered that Test A is too easy and Test B is too difficult. Exactly! But this condition is one kind of invalidity. The expert in measurement would say that such tests do not discriminate differences in ability. This gives us another point of view of the nature of validity, viz., that *valid tests arrange (or rank) pupils on a scale of ability.*

Granting all this, how is a teacher to eliminate items that are functionless because they are passed or failed by all? There are two ways to do this: (*a*) by her judgment, a rough-and-ready but far from valueless method; and (*b*) by giving the test, and by means of tabulations determining which items are too hard or too easy to function. Keep in mind, however, that a few very easy items (passed by 95% to 100%) are desirable for the sake of encouragement, and that

[1]C. C. Weidemann, "How to Construct the True-False Examination," *Teachers College Contributions to Education*, No. 225 (N. Y.: Columbia University, 1926).

a few very difficult items (passed by 0% to 5%) should be placed at the end of the test to give it "top," or sufficient difficulty to prevent perfect scores (non-measurement). Ordinarily, judgment plus later inspection of the test papers will suffice to eliminate most of the functionless material.

The last one of the list of principles for assuring validity calls for some comment. In making a standard test elaborate experimental try-outs are made of the test items in order to determine their degrees of difficulty. The tests are given, scored, and the number of successes (or failures) is tabulated for each item. The easiest one is then placed first, followed by the next easiest, and so on until the most difficult is placed last. At the same time any excess of items passed by all or failed by all is corrected as previously described.

Assume that we have 100 items of varying difficulty with which to make a test. Suppose that we make up this test in two editions, as follows:

Edition I: Items arranged in exact order of difficulty, the easiest ones first; then gradually increasing in difficulty until the last item is the most difficult.

Edition II: Items arranged in the order determined by placing them all in a hat and drawing them one at a time until all are drawn (strictly chance order).

Figure 1 illustrates graphically the two editions of this hypothetical test. Each vertical line represents a test item. The height of the vertical bars indicates the difficulty of the items (in terms of per cents of failures). A short line means an easy item; a long line a very difficult one.

Distribute Editions I and II in chance order to a class of pupils so that half receive I and half receive II. Now it might readily happen that a very difficult item (e. g., Item No. 4) occurs as one of the very first items in Edition II. Those pupils who received Edition II would "hang up" on this hard item; i. e., they would waste time on it, and they

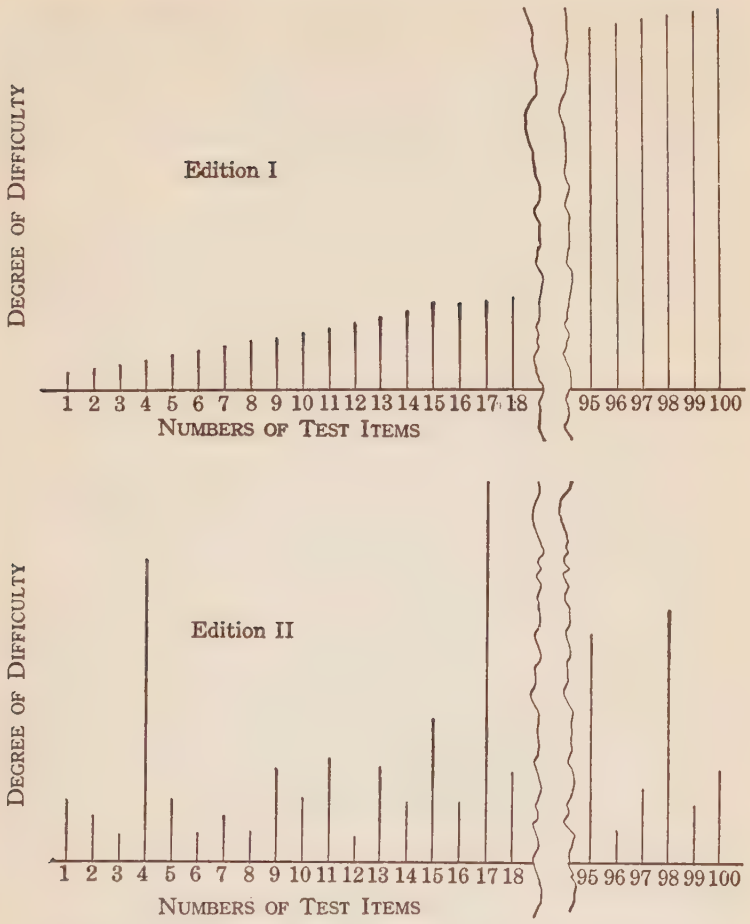

Degree of Difficulty
Edition I
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
95 96 97 98 99 100
Numbers of Test Items
Degree of Difficulty
Edition II
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
95 96 97 98 99 100
Numbers of Test Items

Fig. 1.—Illustrating the arrangement of test items in increasing order of difficulty (Edition I) and in chance order (Edition II). The length of the vertical bars indicates the degree of difficulty.

might fail to answer it even after many minutes of thought. Had they skipped it or had that item been at the end of the

test where it belongs, they might have answered a dozen easier items which were further along in the test while they were puzzling over this one difficult point. Progress through such a test would be by jerks, a run of easy items answered in a few seconds, then perhaps minutes of delay, then more rapid progress, a second long delay, and so on.

The pupils receiving Edition I would move along regularly through the easy items at the beginning of the test. The items would become gradually and almost imperceptibly more difficult. They would finally work, with few delays, into the level of the test where they would be stopped by the limit of their abilities. The time allowed would be better distributed because they met the easy items first, answered these rapidly, and had most of their time left to think about the puzzling and difficult items.

This rough picture of the differences brought about by good and bad arrangements of test items shows one of the reasons why well-made standard tests are superior to most other forms of examinations.

It is not intended that teachers should gain the impression that all of their tests should be subjected to the laborious and expensive experimentation necessary to the correct arrangement of items in order of difficulty. Present discussions are directed at laying the basis for real insight into and understanding of test validation methods. The arrangement of test items in order of difficulty can be done reasonably well by the judgment of teachers. If two or more teachers can co-operate in obtaining such judgments, the average rating for difficulty is almost certain to be considerably better than a single teacher's judgment.[1]

[1]In long examinations test items are often grouped by topics, each major division of subject-matter constituting a separate *part* of the test. This is often done to facilitate diagnosis. In such an event the arrangement of items from easy to difficult may be carried out within each part.

In this connection the author feels it necessary to remind the test maker that if such separate parts are to yield diagnostic values, each part must be made long enough to form a highly reliable test standing alone. In other words, if a test is to yield four separate marks or scores on a given pupil, then each part must be made as reliable as would ordinarily be necessary for an *entire* examination.

An experimental method of validating individual test items. In case a teacher wishes to validate test items with a greater degree of refinement than the method of individual judgments or average of several judgments, there is a reasonably simple method which is quite serviceable. The steps are as follows:

1. Make up the test items, arranging them by inspection in order of difficulty.
2. Give the test to the class, allowing time for all to attempt every item.
3. Score the test.
4. Arrange the papers in order of size of scores.
5. Find the median mark and separate the papers into two classes: (*a*) those above the median, and (*b*) those below the median. Call the first group the "good" pupils and the second group the "poor" pupils.
6. Tabulate the number of pupils passing (or failing) each individual test item, keeping separate tabulations for the "good" and "poor" groups. Express the passes (or failures) in per cents.
7. Study the per cents for "good" and "poor" groups. Reject items where the "poor" group shows percentages of successes as high as or higher than the "good" group. Such items do not differentiate abilities. The best items will show the largest differences in successes in favor of the "good" group.

We can illustrate this method by an example. Suppose we have carried out the above seven steps and obtain results like those of Table 1 on page 38.

Item 1 is neither too easy nor too difficult. In this respect it is satisfactory. But it does not differentiate between high abilities and low abilities. This item does not injure the test, but it might be replaced to advantage by an item showing greater differentiation.

Compare Item 1 with Item 2. They show the same difficulty when both groups are considered. Item 2 is greatly superior to Item 1 because it discriminates between pupils of high levels of ability and those of low levels of ability. Items like No. 2 will make a more valid test than items like No. 1.

TABLE 1

PER CENTS OF "GOOD" AND "POOR" PUPILS ANSWERING INDIVIDUAL ITEMS OF A TEST

ITEM	PER CENT OF CORRECT ANSWERS		
	"Good" Group	"Poor" Group	Both Groups
1	14	14	14
2	21	7	14
3	0	6	3
4	84	16	50
5	53	49	51
6	100	98	99
7	0	0	0
8	100	100	100
9	0	8	4
10	50	50	50
Etc	..	..	..

Item 3 should be discarded. It is rather difficult, and backward pupils do better on it than really superior pupils. To throw this item out will increase the validity of the test. The same comments apply to Item 9.

Item 4 is a good one. It discriminates sharply between good and poor pupils. It is well within the abilities of both groups.

Item 5 has about the same *average* difficulty as Item 4, but it is greatly inferior because low-grade pupils do almost as well on it as high-grade pupils. Like Item 1 it does not hurt the test, but replacing it with one like Item 4 will raise the validity.

Item 6 is very easy. A few such items may be kept in order to encourage the pupils. If so, they should form the first items of the test. Only a few such should be kept as they are too easy to help much in measuring, and "poor" pupils do almost as well as "good" on such tasks.

Items like No. 7 should be thrown away except for a few to give "top" to the test, i. e., to prevent perfect scores.

Item 8 is similar to Item 6, although even less difficult.

Item 9 is too hard and also does not discriminate well between different levels of ability. It probably should be discarded.

Item 10 suffers from the same fault as Item 5. It is by no means as valuable as Item 4, which is of about the same difficulty. It may be retained, although better items can be found.

As has been stated, the method of experimental validation of individual items is usually employed only when standard tests are under construction. Certain informal classroom tests may justify the effort expended in making a study of the values of individual items. A defensible procedure has been given to cover such cases. In most cases it will not be necessary to resort to such elaborate methods. The judgments of one or more teachers will usually accomplish a degree of refinement commensurate with the needs of the average test. One more thing should be kept constantly in mind; viz., *long tests may be expected to be more valid than short tests, and if a test is made long enough, it will usually yield a reasonably valid measure even if many individual items are faulty or worthless.*

The preceding pages have served as an introduction to the main concept of validity. We will return to this topic after the meaning of reliability has been considered, and again in Part II where we consider the actual plan for building valid and reliable tests.

Summary of concept of validity. The following summarizing statements may serve to bring together and clinch the various ideas advanced as relating to validity:

1. Validity is the degree to which an examination measures what it is claimed to measure.
2. Validity includes reliability as well.
3. Validity may be defined as the degree to which the test parallels the actual flow of instruction, and of the care exercised in choosing important materials, in excluding non-essentials, and in producing a correct balance of the materials used in the examination.
4. In the usual school examination, validation rests largely on competent opinion. Standard tests, on the other hand, are validated, in part, by controlled experimental methods.
5. Validation of tests and examinations will be facilitated by building a skeleton of "Table of Specifications" before actual work is begun on the test. (See Chapter VII.)
6. Test items passed by all or failed by all have no validity.
7. The validity of an examination is increased by arranging the items in order of increasing difficulty.
8. Validity of test items is reduced when the statements are ambiguous, faulty in sentence structure, or otherwise not clear.
9. The greater the difference between the percentages of successes of strong and weak pupils on a test item, the more valid (discriminating) is the item.
10. Long tests tend to be more valid than short ones.

RELIABILITY

Definition of reliability. Reliability is second only to validity as a criterion of the worth of a test or examination. We might say that the second most important fact which we can know about a test is the degree of reliability which it possesses. As in the case of validity, a number of statements

are given below, which, collectively and individually, serve to define the concept.

1. Reliability refers to the degree to which a test measures whatever it does measure; not necessarily what it is claimed to measure.

2. Reliability refers to the degree of accuracy of measurement.

3. Reliability refers to the amount of confidence that may be placed in the mark or score on a test as a measure of some ability of a pupil.

4. Reliability is one aspect of validity. A valid test is necessarily reliable, but a reliable test need not have high validity, or for that matter have any validity at all for a particular purpose.

5. Reliability refers to the stability of an estimate of a pupil's ability from one sampling to another. For example, if a certain standard test is marketed with six equal or equivalent forms, this test is not reliable if the fluctuations of pupils' scores from one form to the next is very large.

6. If a valid and reliable test in one subject (say, geography) is given and the results are labeled as of another subject (say, manual training), the actual scores may remain reliable but be entirely invalidated by such misnaming. Thus a thermometer used to measure the velocity of the wind will retain whatever degree of accuracy (reliability) of measurement was built into the instrument by the manufacturer, but the readings, however accurate, will be *invalid* measures of wind velocity. We see that in one sense validity is more specific than reliability; i. e., misuse of a test may injure its validity more than its reliability.

Reliability is thus seen to be a more restricted term than validity; it is one aspect of validity. Validity implies reliability, but the converse is not necessarily true. This point is often confusing, but later sections of this volume may help to clarify these relationships.

Principal methods of insuring reliability to tests. There is little to be added to the discussion already given in the sections dealing with validity. Validity and reliability are so closely related, so far as actual test construction is concerned, that provision for the former in large measure takes care of both. The best approach to the technique of insuring reliability in tests and examinations is a consideration of the two principal means of guaranteeing reliability, viz.:

1. Objectivity of scoring or evaluating.
2. Character of the sampling included in the test items.

Objectivity as a means toward reliability. The objections to the traditional examination have very largely been centered about the variations which are bound to occur when equally competent teachers mark the same examination papers. Such variations obviously tell nothing about the merit of the paper.

They constitute on the other hand a fertile source of unreliability. We have termed such unreliability, *unreliability due to subjectivity*. In contrast to questions open to personal differences in the marking, there are many types of questions or items which do not permit even the slightest differences of opinion in deciding whether the answer is correct or incorrect. Such test items are termed *objective*. Examples of objective test methods are the familiar true-false, multiple-choice, matching tests, etc. Less perfectly objective are the completion tests, which, with care, may be so phrased as to be as objective as practical requirements suggest.

Example A shows a subjective type of question. Example B, in contrast, is an almost purely objective arrangement of the same test material. Example A was one of five questions in an examination in physiology, twenty points being allowed for a perfect answer. Example B provides twenty blanks, each one correctly completed to give one point of credit.

EXAMPLE A

Trace the complete process of the digestion of food from the time it enters the mouth until the waste products are eliminated.

EXAMPLE B

The mouth is concerned with digestion in two ways: first, the grinding action of the ....., and second, the chemical action of the enzyme ....., which acts on ..... changing them into ..... In the stomach the most important enzyme is ....., which starts the digestion of the ..... The gastric juice also contains an ....., which helps to kill the bacteria causing fermentation. The small intestine, which is a coiled tube about ..... feet in length, secretes a digestive juice itself as well as receiving the juices from the ..... and the ..... The bile aids chiefly in the digestion of ..... and in destroying acids from the stomach. The most important digestive juice in many ways is that of the ....., which contains three important enzymes which act on ....., ..., and ..... The absorption of the digested foods is aided by the many finger-like projections in the ..... intestine known as ..... The waste materials collect in the .... and pass on out. If this waste material is not rapidly cleared out, ... are formed which cause disease.

Example B is not entirely objective, but study of the possible insertions on each blank in turn will show that there are very few answers of merit which could be written in on any blank. If 100 teachers should grade a given pupil's answers to this exercise it is to be doubted whether the greatest difference among these hundred teachers would be as much as five points. Moreover, if a set of scoring rules were drawn up and adhered to in the marking, the variation among 100 teachers might be reduced to one or two points at most. In a long examination, the variation represented by one or two points in twenty is of slight moment.

In Chapter III we shall see that questions similar to Example A often show disagreements as great as from three to twenty points when 100 teachers grade the same answer.

We could eliminate all subjectivity in grading the question under discussion by making a true-false, multiple-choice, or simple recall test covering the same points. Example C below shows a twenty-item true-false test covering substantially the same ground as Examples A and B.

EXAMPLE C

1. The principal function of the teeth in digestion is to grind up the food.	True	False[1]
2. The salivary juice contains an enzyme called pepsin.	True	False
3. The enzyme of the saliva acts on starches.	True	False
4. Ptyalin changes sugars into starches.	True	False
5. The gastric juice contains an enzyme known as ptyalin.	True	False
6. The gastric juice starts the digestion of protein.	True	False
7. The gastric juice contains hydrochloric acid.	True	False
8. The acid in the stomach helps to destroy bacteria causing fermentation.	True	False
9. The small intestine is about six feet long.	True	False
10. The small intestine receives three principal digestive juices.	True	False
11. The bile digests mainly carbohydrates.	True	False
12. The most important digestive juice is that from the liver.	True	False
13. Pancreatic juice contains three important enzymes, acting, respectively, on starches, protein, and fats.	True	False
14. The protein-digesting enzyme of the pancreas is called amylopsin.	True	False
15. Lipase acts on fats.	True	False
16. The finger-like processes on the walls of the small intestine are called cilia.	True	False
17. The villi are absorptive organs.	True	False
18. Undigested materials are stored in the large intestine until needed.	True	False
19. Important digestive enzymes are formed in the walls of the large intestine.	True	False
20. If the large intestine fails to act, toxins are formed which may cause disease.	True	False

[1]The practice of printing the words "true" and "false" at the right of each statement is not the best method of arranging true-false tests except, perhaps, with very young pupils. Later sections of this volume show other plans which permit much more rapid scoring. For example, the signs + and − or + and 0 may be used to indicate true and false statements respectively.

Example C is perfectly objective and may be scored by a clerk entirely ignorant of physiology with quite accurate results provided a scoring key is furnished and reasonable care is exercised to avoid mistakes of carelessness. Example C may not prove to be more reliable than Example B, but both B and C are almost certain to be superior to Example A. Example B is not quite objective, but this limitation may be of less weight than is the unreliability introduced into the true-false test (Example C) by the opportunity for guessing. The exact merits of these three forms of what is substantially the same subject-matter need not be settled here. The point at issue is that of defining objectivity of scoring, and of showing its relation to reliability.

Chapter III will take up the experimental evidence on the question of unreliability arising from subjectivity in detail.

Reliability as affected by sampling. The term *sampling*, and perhaps to some extent the idea of sampling, is ordinarily not highly conscious in the mind of the teacher in thinking about examinations and examination procedures. This is not altogether true, as teachers recognize that a two-, three-, or even a five-question examination of the traditional sort is not entirely adequate. They prefer that at least ten questions be asked in an examination of any considerable importance. If asked the reason why a two-question examination is not as desirable as a ten-question test, the average teacher will answer, and rightly, "The former is too short." This is equivalent to saying that a two-question examination is not very *reliable*. Reliability is thus a new term for an old, but not thoroughly analyzed, idea. This brings us to the conclusion that examinations should be relatively long. Why? Because a long examination is a more adequate sampling than a short one, all other factors being equal.

A short examination suffers from many defects. Prominent among these are such facts as: (*a*) a short examination does not cover the ground thoroughly; (*b*) a short examination places too much premium on the knowledge of the particular ground covered but tells nothing about other equally important divisions of the subject-matter; and (*c*) a short examination penalizes unduly a pupil who happens to have little knowledge of the particular questions but has otherwise a good knowledge of the subject, or, conversely, a short examination introduces an element of luck in that a pupil might know the particular questions asked but be shaky on others equally important but unasked.

A test or examination is always a sample. Measurement is never complete; it always represents a sample of abilities. Other things being equal, the longer a test, the more adequate the sampling. The more adequate the sampling, the more reliable (and hence indirectly the more valid) the test.

Since testing is sampling, any test score involves a certain amount of error of measurement. To say that a test is unreliable is synonymous with stating that it has a large error of measurement. Reliability is accuracy of measurement. As a test is made longer and longer, it becomes more and more reliable, provided the test is increased by equally good test items. This idea is not new. Teachers are aware that the practice of basing term marks on a single examination is dangerous (unreliable). The average of two examination marks is safer. The average of many examination grades is still better. We can think of two examinations as one examination doubled. We can regard the average of ten examinations as one examination ten times as long as any one of the ten.

If we accept the point of view that increasing the length of a test raises its reliability, and that this process may be continued without limit (in theory at least), we arrive at a point of view that: *A test infinitely long is perfectly reliable.*

To increase the reliability of a test, ordinarily it is sufficient to lengthen it, i. e., *extend the sampling.*

We can find an analogy to the unreliability of a very short test (narrow sample) in the hypothetical experience of a teacher. Suppose that it is now April or May and the teachers of your school system are about to receive notices of retention or dismissal for the following year. In order to form a basis for recommendations to the board of education your superintendent decides to base his judgment upon an unannounced five-minute visit to each teacher's classroom. Upon the basis of his observation of a teacher's classroom activities for five minutes, he decides to "drop" her. In another classroom he finds a very interesting recitation in progress (perhaps the only good one all week). He recommends that this teacher receive a $200 increase the next year. And so on.

Would teachers feel the justice of this plan? A five-minute visit is fundamentally similar to a five-minute test—it is too short to afford a *reliable* basis for important judgments. But if the observational classroom visits were made frequently enough to sample the general run of the activities of a teacher, in the long run something could be told about the general worth of the teacher. Or, if the superintendent, principal, and other general supervisors confer and exchange notes, the combined judgment of several such persons will have more value than that of the superintendent alone.

The position has been taken repeatedly in the preceding pages that *measurement is always limited sampling,* but it is only fair to state that at least one writer on educational measurement, Dr. C. W. Odell, challenges this view.[1]

[1]C. W. Odell, in reviewing the author's *Improvement of the Written Examination,* in the *Journal of Educational Research,* December, 1925, p. 42, says: "... Ruch states that 'measurement is always sampling,' again 'that an examination of ten questions, or worse still, five questions ... is a very small sampling is evident.' Although both statements express general truths, the reviewer does not believe that they always hold. In a limited field one can secure a complete measure of ability. For example, a test may include all the addition combinations of simple digits. Likewise ten or even five questions may give more than a very small sampling if they are topical or in some other way call for large amounts of material."

One instance raised as a criticism of the view that measurement is always limited sampling is the possibility of complete measurement of a narrow ability or function like the 100 addition facts. Let us consider this case, as it is a common school situation and an excellent illustration of the differences of opinion, if any, between Odell and the author.

Complete vs. incomplete measurement. There are but 100 basic addition facts when we include the zero combinations and call both orders of addition of any two digits (e. g., 6+4, and 4+6) different facts, which modern authorities agree upon as necessary.

The author cares little about who is right or wrong so far as the controversy is concerned, but the illustration is an excellent one for bringing out certain concepts related to sampling and unreliability of measurement. The reader may choose between the two views at will, but the difference of opinion bears upon our present definition of unreliability.

The following simple experiment was carried out as an illustration:[1]

1. The one hundred basic addition facts were placed on one hundred small pieces of cardboard of uniform size.
2. The cardboards were then shuffled thoroughly.
3. They were then drawn, one at a time, and placed in the order of drawing as a test. This was called Form A.
4. The shuffling and drawing process was repeated four times to yield Forms B, C, D, and E.
5. The five forms (A to E) were mimeographed and administered to two groups of pupils on five successive days.

NOTE: Class X was a beginning class which had not as yet thoroughly mastered the addition facts. Class Y was a strong third-grade class.

6. The scores obtained were tabulated. This tabulation is given here, all incomplete sets being thrown out.

[1]By Miss Celia Gifford, Sunshine School, Berkeley, California.

It should be noted that all five examinations contained exactly the same combinations (*all the combinations*), the order only being different from form to form. Each of the five forms was therefore a "complete sampling" in the sense used by Dr. Odell.

Pupil	Form A	Form B	Form C	Form D	Form E
			Class X		
1	99	100	100	100	99
2	86	98	100	100	100
3	89	100	99	100	99
4	100	98	100	100	100
5	100	95	100	100	100
6	100	100	99	98	96
7	87	98	99	100	100
8	77	100	100	99	100
9	78	32	94	57	76
10	90	98	93	80	99
11	18	90	39	38	47
12	79	79	90	91	91
			Class Y		
13	100	100	100	97	99
14	94	98	93	94	93
15	97	99	97	98	100
16	100	100	92	100	100
17	98	100	100	98	98
18	93	100	98	97	98
19	99	100	99	99	99
20	87	93	90	87	90
21	99	100	99	99	99
22	99	96	93	96	92
23	99	99	100	100	100
24	98	99	100	100	100
25	88	78	81	76	89
26	99	100	100	100	100
27	89	90	94	92	86
28	100	98	100	100	98
29	100	100	98	100	95
30	100	100	99	100	100

There are several instances of perfect agreement on two, three, or four forms. *There was not a single case where a pupil*

earned exactly the same scores on all five forms. Pupils 26 and 30 made but one error each in 500 attempts. These single errors are unquestionably the results of temporary "slips" or confusion. Pupil 16 was consistent except on Form C where he "fell down" noticeably. Pupils 9 and 11 are unaccountably erratic. We can but wonder what the causes were. Differences of five or more points between any two forms are very common, e. g., Pupils 2, 3, 5, 7, 8, 9, 10, 11, 12, 14, 16, 18, 20, 22, 25, 27, and 29.

It is to be expected that Class X would show larger fluctuations than would Class Y since the former were beginners in arithmetic and the latter had had more than a year of instruction. The magnitude of the disagreements, at times, comes as a surprise to anyone who has not actually carried out the equivalent of our procedure.

We are now ready to generalize. The concept of sampling is dual in character. It includes:

1. Limited sampling of the actual subject-matter. It *is* possible in the case of narrow functions to eliminate this source of unreliability (as in the case of the present illustration).

2. Limited sampling or unreliability arising from psychological factors inherent in the mind of the pupil under examination. Such psychological factors include carelessness, undue haste, lapses of attention, state of health, fluctuations in effort, variations in motivation, etc.

The more practical point of view would seem to be that *sampling is never complete until fluctuations in performance (the pupils' scores) are completely stabilized, even if the subject-matter is completely covered.* Stabilization was not completed in our experiment even after five administrations of the same (?) test, although for all practical purposes the average score on the five forms would be a far more reliable measure than we can ordinarily expect to obtain in school work.

THE MARY E. BRYAN LIBRARY
THE SOUTHWESTERN BAPTIST SEMINARY

The criticism really reduces to a matter of definition. Dr. Odell seems to prefer the narrower point of view; the author has insisted on the dual view defined above.[1] The main point of interest in the present discussion attaches to the isolation and emphasizing of the fact that fluctuations of performance arise from psychological factors, and that these variations enter into the test scores and hence cause unreliability of a single measurement.

It is to be doubted whether it is possible to construct five or ten questions which will sample even half of the field ordinarily covered by a major examination. To be sure, one might make up an examination in physiology somewhat as follows:

I. Describe fully the cell basis of the human body.
II. Discuss the complete process of digestion.
III. Name the principal bones, tell how the skeleton is articulated, and describe its functions.
IV. Give in detail the facts about respiration, the respiratory organs, the relation of respiration to circulation and tissue-building, etc.
V. Etc.

Since physiology texts are often organized around from ten to twenty main topics like cell structure, the skeleton, circulation, respiration, the nervous system, germ diseases, etc., it would be theoretically possible to "cover" the entire field in ten or twenty questions. But in what sense can it be said that the subject has been "covered" (sampled completely)? The main topics are all in to be sure. But will the pupil write everything that is important about all ten or twenty? Will he write a half, a third, a fifth, or a tenth, or less of what he actually knows about digestion? The general nature of the answers to these questions must be evident by now. The following pages report some hitherto unpublished experimental evidence on this point.

[1]After this manuscript had gone to press, Dr. Odell's splendid treatment of the objective examination was published under the title, *Traditional Examinations and New-Type Tests*. From a discussion appearing therein (p. 42) it seems that Odell and the present writer are now in substantial agreement on the issue of "complete" sampling.

27203
371.271
THE CINCINNATI BIBLE SEMINARY
LIBRARY

Talbott's study of the sampling afforded by essay examinations.[1] E. O. Talbott, working under the direction of the author, has carried out a study of the traditional examination from the standpoint of sampling. His specific problem is: what fraction of a pupil's knowledge is elicited when he is asked to "Discuss fully" (or some equivalent phrasing) a given topic?

To approach a rough answer Talbott proceeded as follows:

1. He gave an essay examination of from three to ten broad discussion questions.
2. The pupils wrote their answers, unlimited time being given, although accurate records of actual working times were kept.
3. The pupils took immediately a long objective test on each topic represented by an essay question in the first examination. Working time was again recorded. (The order of essay and objective tests was alternated from one experiment to the next.)
4. The essay questions were scored for the number of *ideas* or *facts* written by the pupil.
5. The objective tests were scored in such a way as to correct for chance or guessed successes.
6. The ratio between essay scores and objective scores was computed for each pupil, and similarly for working time ratios.

This is of course a very rough procedure, although one which is far from valueless in forming general notions about the merits of old- and new-type examinations as devices for sampling completely the knowledge or skills possessed by pupils. The errors of the method are likely to be quite conservative since it is obvious that no objective test could cover everything that a pupil might know. If the essay papers are scored generously, as in Talbott's study, the ratios are likely to be smaller, not larger, than the truth.

[1]Unpublished. To appear in the *Journal of Educational Research.*

Table 2 on page 54 presents a summary of the findings. The first line of entries refers to an examination in eighth-grade geography. It will help to describe Talbott's procedure to discuss this geography examination briefly.

The essay test consisted of the following questions:

I. Discuss the continent of Europe.
II. Discuss the green Northlands.
III. Discuss fully Ireland, Scotland, and England.

The objective test covering the same ground included 247 items, 201 being true-false and 46 simple completion exercises. Samples of these items follow:

1. Europe is next to the smallest inhabited continent. True False
2. Europeans and those who have gone recently from Europe rule most of the world. True False
3. Europeans have settled North and South America and Australia. True False
4. The numerous lakes of Ireland and Scotland were probably caused by .....

The completion tests were scored simply "number correct," but the true-false were scored "rights-minus-wrongs" in order to allow for chance or guessed successes.

Talbott's method of scoring the essay questions may be shown by a portion of an actual paper. He read each paper carefully, drawing vertical lines at the end of each separate thought or fact to the best of his judgment, attempting to be generous and giving no penalties for errors of spelling, grammar, etc. No credit was given for incorrect statements. Thus:

Europe is just east of U. S. | There are many manaured (manufactured?) things sent out | and lots of imports. | Some of this country of Europe is ruled by a King | and others just about the same as U. S. | Etc.

Five credits were given for this portion, although the first statement is somewhat questionable.

We are now ready to examine Talbott's findings.

The row of averages at the bottom of Table 2 is best for purposes of drawing probable conclusions. It seems likely that the essay test calls forth less than half (.44) of the pupil's real knowledge of the subject. To elicit this two-fifths or half of a pupil's knowledge required two times as much working time (2.01) as was needed for the long objective test. It thus appears that the objective tests used were four or five times as efficient as the essay questions as devices for sampling when we take into account both average working times and sampling ratios. There can be little doubt that the objective test is the more efficient sampling method per unit of working time.

TABLE 2

SELECTED RESULTS FROM TALBOTT'S STUDY OF THE ADEQUACY OF THE TRADITIONAL EXAMINATION AS A DEVICE FOR SAMPLING

SUBJECT	SAMPLING RATIOS $\left(\frac{E}{O}\right)$			TIME RATIOS $\left(\frac{T_e}{T_o}\right)$			N
	Highest	Lowest	Average	Highest	Lowest	Average	
Geography (Grade 8)	.98	.28	.66	2.76	.61	1.31	17
History (Grade 8)	.90	.21	.44	7.30	1.13	3.00	17
U. S. History I	.84	.37	.51	1.91	.93	1.53	14
U. S. History II	.69	.21	.36	2.40	.91	1.85	20
Chemistry I	.81	.24	.41	2.76	.76	1.90	15
Chemistry II	.64	.22	.37	3.20	.94	1.96	20
Citizenship I	.92	.25	.41	5.18	.77	2.03	18
Citizenship II	.68	.24	.38	2.36	.97	1.73	19
General Science I	.72	.22	.47	4.07	1.11	3.05	13
General Science II	.56	.23	.40	2.16	.96	1.70	14
Averages of Columns	.77	.25	.44	3.41	.91	2.01	(Total) 167

Table 2 shows that in certain individual cases ("Highest Column"), pupils did write from two-thirds to practically all that they knew under the stimulus "discuss fully." At the same time the "Lowest" ratios fell as low as one-fifth (.21) to less than two-fifths (.37). In the latter cases the directions to "tell all you know" guaranteed very little.

Talbott's results must not be taken as final, especially in view of the obvious crudities of the method. Thus far he has studied twenty elementary and high-school classes in two school systems with consistent results. More work is needed on this and similar problems before reaching final decisions. It is somewhat surprising that so little critical work has been done on so many of the claims of the traditional examination.

Unreliability usually due to both subjectivity and limited sampling. Unreliability has been treated as falling into two categories: (*a*) unreliability due to subjectivity and (*b*) unreliability due to limited sampling. These two sorts of unreliability are hard to differentiate in many kinds of examinations. They usually are found to operate together in the traditional examination. In standard tests and objective classroom tests, subjectivity may disappear completely, but limited sampling will remain as a disturbing factor. We may consider four sorts of tests by way of illustration:

TEST A: 10 discussion (essay) questions on United States history.
TEST B: 1000 discussion questions on human physiology.
TEST C: 10 true-false questions on grammar.
TEST D: 250 multiple-choice questions on geography.

Test A is a very common type. It is limited in scope since but ten questions are employed. These are discussional in character and hence open to differences of opinion in scoring. Such a test is ordinarily but moderately reliable. It is subject to *both* sources of unreliability, limited sampling and subjectivity of marking.

Test B is hardly a practical illustration. Such a test would take many days. It would be comprehensive, and for all practical purposes limited sampling would not enter. Subjectivity of scoring would remain.

Test C is objective. If 1000 teachers should mark the same pupil's paper, the agreement would be perfect except for occasional errors of carelessness. But it would be a very inadequate sample of a pupil's achievement in grammar.

Test D is both highly objective (probably perfectly so) and a reasonably wide sample. It might be the best test in the lot. Test B, in theory, might be better, but it would be impossible to administer such a test in practice. If well made, Test D is a reasonable approximation to the elimination of both sources of unreliability. Test D might be given in sixty to ninety minutes, a justifiable expenditure of time.

Objectivity is attainable by a change, where possible, to the new-type examination. Subjectivity cannot be eliminated from the essay-type test by any method within the limits of practicability. Two solutions have been proposed for remedying the unreliability of the traditional examination without its abandonment: (1) having a number of teachers grade the same papers and taking the average as the mark; (2) using a written set of rules for scoring all doubtful situations. The first mentioned remedy is undeniably efficacious. The difficulty is that it might take dozens of teachers to reduce the personal judgments to a stable basis or average. This method will not work out in practice, as is evident. The second method is somewhat more promising. It will be considered further in Chapter III. The experimental evidence to date makes it questionable whether the refinement possible in this direction will be sufficient to eliminate the main force of the objection to subjective examinations.

A sampling theory of examinations.[1] "Examination practices at the present time make use of two more or less distinct theories of sampling. The first of these may be called the *intensive* sampling and is represented by the traditional essay type of examination. The second, or

[1]Quoted from G. M. Ruch *et al, Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Co., 1926), pp. 12-14.

extensive sampling, is characteristic of the more recent new-type or objective examinations. The former examination usually consists of five or ten questions which are to be answered exhaustively. The latter is more likely to comprise from 50 to 250 or more narrow questions which sample widely but not intensively.

"It is not always borne in mind that any examination is at best a limited sampling of the total field which might be covered by the examination. Testing is therefore invariably partial, never complete. Unreliability due to sampling can be reduced by increasing the number of questions asked. Theoretically, an examination is perfectly reliable only when the sampling is infinitely long.

"The two theories of sampling as applied to examinations may be illustrated by the following scheme:

"Let *A*, *B*, *C*, *D*, . . . *N* below represent the topics in a particular school subject. Each should be suitable for one broad question of the usual type. Let 1, 2, 3, . . . *n*, represent single facts or items of information, etc., falling under each of the topics denoted by capital letters. Thus:

A	*B*	*C*	*D*	.	.	.	*N*
1	1	1	1	.	.	.	1
2	2	2	2	.	.	.	2
3	3	3	3	.	.	.	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
n	*n*	*n*	*n*	*n*	*n*	*n*	*n*

[The same facts are shown in diagrammatic form in Figs. 2 and 3.]

"It is logical to suppose (and it can be shown experimentally[1]) that knowledge of item or question *A*1 is much

[1]For example, in standard tests, like the *Thorndike-McCall Reading Test*, it can be shown that the items based upon the same reading paragraph are more highly interrelated than are items of two different reading paragraphs. Here the paragraphs are analogous to the capital letters in the scheme, and the questions or items based on the paragraphs are analogous to the 1, 2, 3, etc., falling under each capital letter.

more likely to guarantee knowledge of items *A*2 and *A*3 than it is to guarantee knowledge of items *B*1, *C*6, and *Nn*. This is merely equivalent to saying that the intercorrelations are higher among items of the *same* column than between items drawn from *different* columns.

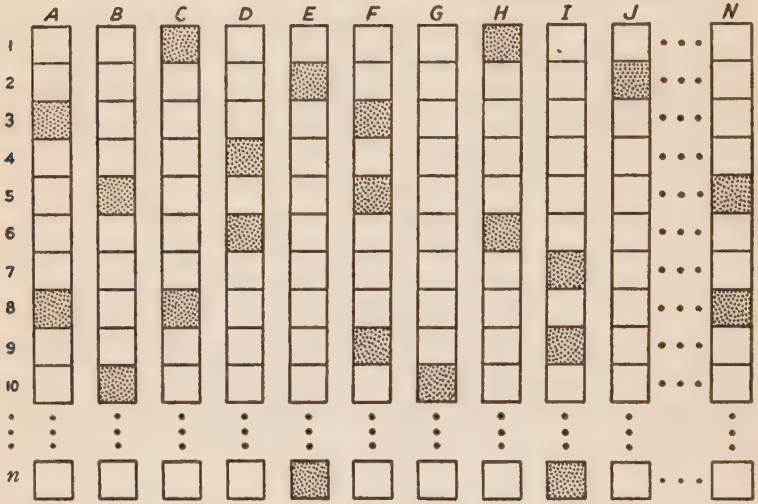


FIG. 2.—Diagrammatic representation of the "extensive" sample. The dotted portions represent the items or facts actually covered by the examination. The numbers and letters are those used in the preceding text.

"The traditional or intensive type of examination tends to the position of including a few (five to ten) columns or topics, these being answered in great detail. The newer objective examination tends to the extensive sampling, i. e., a few narrow items drawn from many columns.

"A priori, the advantage lies with the extensive sampling so far as reliability of sampling is concerned, since such samples are not so greatly affected by occasional faulty questions, the missing of work due to absence from school, and other obvious factors. It might also be pointed out

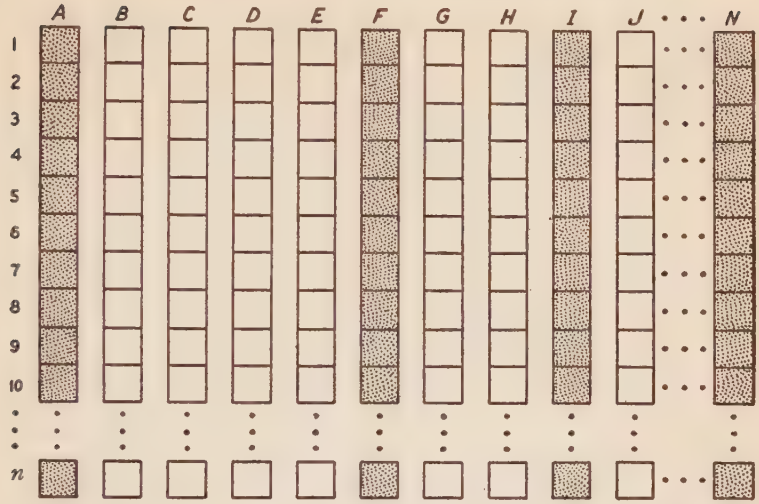


FIG. 3.—Diagrammatic representation of the "intensive" sample. The dotted portions represent the items or facts actually covered by the examination. The numbers and letters are those used in the preceding text.

that the situation with respect to subjectivity of scoring is similar, since the narrow question is less subject to personal opinion than the broader type of question."

The interrelations of validity and reliability. The attempt has been made to approach the concepts of validity and reliability from a number of different angles. Much of the discussion has been somewhat theoretical. This was done intentionally upon the theory that a full understanding of the principles of educational measurement requires the acquisition of a new vocabulary and the analysis of measurement into its principal underlying concepts. The relation between validity and reliability may now be pushed somewhat further in theory.

Figure 4 shows two vertical scales which represent validity and reliability, respectively. Lines are drawn to show the

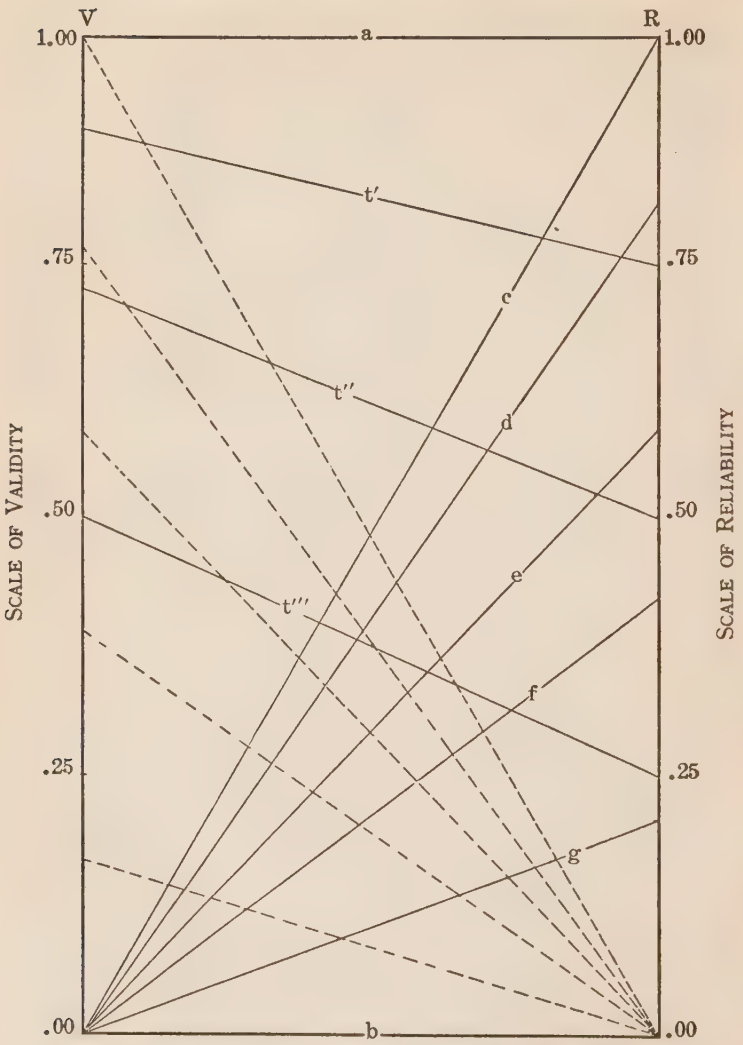


Fig. 4.—Theoretical interrelationship of validity and reliability.

theoretical interrelationships. Solid lines present *possible* relationships. Dotted lines represent *impossible* conditions.

Fig. 4 states certain theoretical relationships between validity and reliability. The scales of validity and reliability are graduated in terms of (correlation) coefficients of validity and reliability.

The following statements are based upon Fig. 4.

1. Line *a* represents a possible situation—a test perfectly valid and also perfectly reliable; in fact, such a test could not be perfectly valid unless also entirely reliable. It need hardly be pointed out that such tests exist only in theory, since all educational measures are both invalid and unreliable in greater or less degree. Line *a* therefore represents a limiting condition.

2. Line *b* is also possible, and is occasionally approximated in actual practice. The test is totally unreliable and hence has no validity.

3. Lines *c*, *d*, *e*, *f*, and *g* have been drawn to show that a test may be highly (in theory, perfectly) reliable and yet have no validity. Such extreme cases do not occur often in actual practice, but such a condition might be found when reliable tests are grossly misapplied.

4. Lines t', t'', and t''' are inserted to show that it is possible for the validity of a test to be somewhat greater than its reliability, if these are expressed in certain quantitative terms. Such conditions need not concern us here.[1]

The preceding discussion is probably sufficient for present purposes. The issues raised will gradually become more meaningful as later chapters present concrete experimental

[1]This statement means that the validity of a test may be as high as the square root of the reliability of the test. In this case validity represents correlation against a perfectly valid criterion. See T. L. Kelley, *Statistical Method* (N. Y.: The Macmillan Co., 1923), pp. 205-208. Kelley gives a formula which is at times of the greatest value in discovering whether it is worth while to attempt to improve the validity of a test through increasing its reliability.

The serious student of educational measurement may perhaps be curious as to the reason why lines t', t'', and t''' were inserted, in which case see the above reference.

studies related to reliability. The treatment of reliability may be closed by a brief summary of the points of view presented. Chapter XV treats of the statistical determination of reliability by means of coefficients of correlation.

Summary of concept of reliability. The following summarizing statements will review the principal ideas centering about reliability:

1. Reliability is second only to validity in constructing educational measurements.
2. Reliability is that phase of validity which refers to the accuracy of a test as a measuring device.
3. Reliability is presupposed when validity is established.
4. Reliability is the stability of numerical scores for the same individual or individuals when equally difficult and similar examinations are applied in sequence.
5. Reliability may not be reduced as greatly as is validity when a test is grossly misapplied.
6. Reliability is guaranteed in two principal ways: (*a*) objectivity of scoring, and (*b*) extended sampling (length of test).
7. Traditional examinations are seldom highly reliable because personal opinion enters largely into the evaluating of the papers.
8. Unreliability due to subjectivity can be nearly or entirely eliminated by the use of new-type questions like true-false, multiple-choice, completion, matching tests, etc.
9. Unreliability due to limited sampling is never entirely eliminated.
10. Measurement is never complete; it always represents a sampling of abilities. (Even in the case of narrow functions, where *all* of the subject can be included in the test, measurement still is sampling, on account of the fluctuations in psychological factors involved in answering the test.)
11. Any test score involves greater or less error.

12. As long as repetitions of the same test (or equivalent tests) show fluctuations in the scores of a pupil, measurement can not be said to be complete, i. e., stabilized.

13. Traditional examinations and new-type objective tests differ in their underlying theories of sampling; the former have been termed *intensive* samples and the latter *extensive* samples.

14. Other things being equal, intensive sampling tends to be the more reliable.

15. Theoretically, a test must be infinitely long in order to be perfectly reliable.

EASE OF ADMINISTRATION AND SCORING

Need for adequate instructions. Objective tests are less familiar to many pupils than are the traditional types of examinations. The test exercises present much more complicated features than the usual essay-type question. Some experience and thought are necessary for the concise and skillful statement of test instructions so that the dullest pupil cannot fail to know what he is expected to do. The teacher must remember that an educational test is designed to measure achievement, *not the understanding of instructions*. It is true that tests calling for the following of directions are sometimes employed in intelligence testing, but such exercises are exactly what must be avoided in educational measurements if valid and reliable results are to be obtained. It is never entirely wise to eliminate directions to the pupils from tests even if the same types of tests have been employed repeatedly. Pupils forget certain cautions and directions between times even if they think they remember perfectly. The few lines of space required for a statement of instructions for a test can hardly be regarded as wasted space.

A few suggestions are given here for the guidance of teachers inexperienced in constructing objective tests.

1. The instructions should tell the pupil *in simple language* what he is to do. They should cover such points as (*a*) what mark he should use to designate his answer (+, −, √, ×, underlining, etc.), where to place the answer, what to do to correct an answer, etc.

2. Avoid difficult terms like "encircle," "underline," "underscore," etc., in favor of "draw a circle (or ring) around," "draw a line under," etc. Words like "encircle" may be used with high-school pupils, perhaps, but certainly not with grammar-school pupils. Even in the case of older children there is no reason to risk the use of a term which might not be understood.

3. Give two or three samples, or even more when a new type of test is employed. It is a good plan to have one or two marked correctly, and then to mark the rest of the samples in unison as the teacher directs.

4. If a new test technique for which it is somewhat difficult to phrase concise directions (e. g., matching tests of many sorts) is used, give a preliminary test or fore-exercise to familiarize the pupils with the mechanics of the test. If this is not done, discount the results from the first administration of such a test.

5. When objective tests are given the first few times, the teacher will do well to circulate about the room watching for pupils who have failed to understand the directions. These should be given enough individual help to get them started to work properly.

6. Instruct the pupils what to do in case they are in doubt about a particular item, i. e., whether to leave it out or to guess at the answer. (See the evidence on this point in Chapter XII.)

7. Inform the pupils whether they are to work as rapidly as possible or to take time to be sure that each answer is correct before proceeding to the next item. (Chapter VII presents a number of adequate sets of directions to pupils.)

Need for economical scoring. The objective test is admittedly time-consuming in its construction. Some of this time can be saved in the scoring. The traditional test is devised rather quickly, but it requires much time for scoring. In view of the extra demands upon the teacher's time, the new-type test should be arranged mechanically, before mimeographing, so that it can be scored economically by means of answer keys or stencils. It is possible to use inexperienced clerical help or even older pupils for scoring objective tests if care has been taken to make the scoring simple.

In Chapter VII a number of devices will be presented which will be useful in facilitating scoring when large numbers of tests must be handled. For the present a few suggestions must suffice.

1. Checking, underlining, encircling, etc., are more economical of scoring time than are words written in. (In completion tests it is usually necessary to have the pupil write the words which have been left out.)

2. It is best to arrange the test so that the responses fall in vertical columns down the page. One column is better than two or more. When the responses form vertical columns, strips of cardboard bearing the correct answers may be placed alongside (preferably at the left) of the pupils' answers. This makes a rapidly scored test. Matching tests lend themselves nicely to this arrangement. Simple recall tests may be made to do so by aligning the terminal blanks in a column. Multiple-choice tests are not so convenient, as the responses will occur in irregular positions on the page. The same is true of completion tests.

3. Multiple-response (multiple-choice) tests can be scored more rapidly if the method of responding is by *number* instead of underlining. This method must not be used with young pupils as there are great dangers of confusion, and directions are difficult of phrasing.

NORMS OR STANDARDS FOR EVALUATING TEST SCORES

Norms not essential. Norms have certain values in connection with the interpretation of standard tests. When, however, the teacher constructs her own tests, there are no available norms or standards of attainment. Norms would add certain facts to the interpretation of objective test results, but the expense of deriving such standards would not justify the attempt.

As a matter of fact, the value of norms has been badly over-estimated, even in the case of standard tests.[1] Carefully derived norms have unquestionable value, but local conditions relative to the course of study, ages of children in the different grades, differences in racial and economic background, variations in mental ability, etc., make general or "blanket" norms uncertain business.

The constructor of objective tests must seek other means of interpretation than through the use of norms. Local norms may be derived with the accumulation of records, and in the long run, interpretations may be made quite as accurate as practical demands suggest.

DUPLICATE OR EQUIVALENT TEST FORMS

Definition of equivalent forms. The term *equivalent* forms has been borrowed from the nomenclature of standard tests. There are several synonyms which are quite common, viz., *comparable* forms, *similar* forms, *duplicate* forms, and *equal* forms; the last-mentioned being a looser expression. When standard tests are prepared, two or more equivalent forms are usually constructed. In general, the larger the number of equivalent forms, the greater the usefulness of the test. The conditions which must be met in making

[1]The reader interested in the justification of this statement to the effect that norms are "over-rated" can find the argument stated in Ruch and Stoddard, *Tests and Measurements in High School Instruction*, pp. 60-62.

several comparable forms of a test are rather rigid if the maximum value is to result from such duplication. We can mention the following conditions as, collectively, defining the term:

1. Equivalent forms must show equal average scores when applied to large numbers of children.
2. The spread of scores (range or variability) should be the same on all forms.
3. There should be no duplication of items from one form to another; i. e., each form should be an independent sampling. However,
4. All forms must sample exactly the same function or ability.
5. The correlation (degree of correspondence) should be as high as possible between forms. (This is the same as saying that each form should be as reliable as possible.)

If all five conditions were met perfectly, and a large number of equivalent forms were given to the same pupils, each pupil would receive exactly the same score on each form, thus:

Pupil	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6...	Form *N*
A	68	68	68	68	68	68...	68
B	106	106	106	106	106	106...	106
C	17	17	17	17	17	17...	17
Etc.	..	..	..	..	..	...	..

Such exactness never occurs as a matter of fact, for many reasons, chiefly unreliability and the gains through such a series due to practice effects.

Values of duplicate forms in objective tests. It must be apparent that a close approximation to the conditions which have been laid down for equivalent or similar forms could only be obtained by elaborate experimentation. As has

been said, the more carefully constructed standard tests are published in from two to five or more forms.

Roughly equivalent forms would serve a number of uses in the case of informal objective tests, were such available without great expenditures of time and energy. Let us consider some of these uses.

1. Every teacher is burdened with make-up examinations for absent or failing pupils. Conditions usually preclude the repetition of the regular examination. An equivalent form of the examination would be valuable. If the teacher makes up a new test, she can hardly avoid the inequalities of difficulty, variability, etc., which we discussed at some length previously.

2. There are always a certain number of doubtful cases in the results from any examination. Pupils occasionally perform very badly on a particular examination in comparison with their general records. Such doubtful cases should be re-tested. However, there is often little gain in giving a second test not known to be comparable to the first.[1]

3. Duplicate forms may be distributed in rotation to pupils taking examinations and thus prevent cheating since no two adjacent pupils receive the same questions.

4. A very common method of teaching is to lay out the work for the class in *units* ("projects," "contracts," etc.). Such an organization is an aid to individualized instruction. A pupil works on one unit until, in his judgment, he is ready to go on to the next. The teacher sometimes tests the pupils individually on the unit supposedly mastered. At times the pupil must return to that unit for further work before he is

[1]It is worth while to digress somewhat at this point and call attention to a phenomenon of test scores, which unfortunately is too little appreciated. It has been proved that very high test scores are more likely to be in error *upward* (i. e., too high), and that very low scores are more likely to be *lower* than the truth; i. e., if a number of strictly equivalent forms are given and the results averaged, the average scores tend to move toward the average of the group. The expression is that *test scores, due to unreliability, regress on the mean* (*average*). This fact was discovered by Sir Francis Galton during a study of the inheritance of stature. Regression is a statistical term for the fact that extreme scores are less reliable than those falling nearer the middle of the class.

It is always well to re-test very high and (especially) very low cases.

permitted to go on. It is a question here of duplicate forms or the less satisfactory plan of repeating the same test one or more times. There is more incentive to the pupil to be allowed a "fresh" examination.

5. The last advantage of duplicate forms is perhaps the most important, as it gives a certain aid in a persistent and perplexing problem in grading school work. The typical practice is to change the examinations from semester to semester or at least to repeat them only at intervals of several years. This practice suffers from a serious limitation, *it effectually blocks the accumulation of records which would generalize and refine our bases for assigning school marks.*

Full discussion of this point is reserved for Chapter XIV which deals in some detail with marks and marking systems. For the present it is sufficient to point out that the usual practice of building examinations anew, year after year, can result only in large differences in the difficulty of such examinations from time to time. Such differences make it almost impossible to tell whether successive classes differ in average ability or whether the *apparent* differences are nothing more than variations in the difficulty of the examinations. Well constructed duplicate examinations are sufficiently equal in difficulty to enable the teacher to compare successive classes rather fairly, and, further, to pool the test scores of such classes, year by year, arriving finally at a reasonably accurate local norm. The use of duplicate forms lends itself to a generalization of experience and an accumulation of records which will greatly refine grading practices. By assuming the duplicate forms to be equal, for practical purposes, all pupils over a period of years are graded on the *same scale* of marking.

CHAPTER III

OBJECTIONS TO THE TRADITIONAL EXAMINATION

The objections stated. We have already seen that the principal objections to the essay-type examination reduce to the question of *subjectivity* of scoring. There are minor objections which have been advanced from time to time, some of these being closely related to the matter of unreliability due to lack of objectivity; but others raise questions of economy, sampling, etc. The commonest objections are:

1. Subjectivity of scoring lowers the reliability.
2. The sampling must be limited to a small number of broad questions.
3. The time required to write lengthy answers is excessive.
4. These examinations encourage bluffing.

Before undertaking to comment on these reputed limitations in detail, it will be best to survey a few selected studies on the general problem of the reliability of school marks and marking systems.

INVESTIGATIONS OF TEACHERS' MARKS

Johnson's investigation. Table 3 is quoted from Kelly's arrangement of Johnson's study of marks in the University of Chicago High School.[1] That the standards of the various departments listed in Table 3 show wide variations needs no comment. English teachers fail almost three times as many pupils as do domestic science teachers and give but half as many *A*'s. A pupil's chance of getting an *A* in German is approximately twice as great as his getting one in French.

[1]F. J. Kelly, "Teachers' Marks," *Teachers College Contributions to Education*, No. 66 (New York: Columbia University, 1914), p. 11.

TABLE 3

THE DISTRIBUTIONS OF THE MARKS OF THE SEVERAL DEPARTMENTS OF THE UNIVERSITY OF CHICAGO HIGH SCHOOL
(From Johnson)

DEPARTMENT	TOTAL NO. OF MARKS	% OF *F*	% OF *D*	% OF *C*	% OF *B*	% OF *A*
Greek and Latin	868	10.6	16.1	31.8	23.5	17.9
German	416	8.4	19.5	26.4	28.6	17.1
French	475	10.9	18.7	33.0	28.0	9.3
English	1514	15.5	21.7	32.8	23.4	6.5
Mathematics	1466	14.5	25.2	27.6	21.1	11.5
History	825	8.1	15.9	31.2	30.0	14.7
Science	672	8.3	16.8	27.7	32.6	14.6
Domestic Science	176	5.7	2.3	27.3	51.7	13.1
Average	(7297)	11.5	18.9	30.6	27.0	12.0

Variations in teachers' markings in a large city high school. Hendrickson[1] has given us another example of the differing standards in marking pupils in the several departments of a city high school. His tabulation follows:

TABLE 4

DISTRIBUTION OF MARKS BY DEPARTMENTS, VAN NUYS, JUNE, 1927

DEPARTMENT	*A* %	*B* %	*C* %	*D* %	*E* %	TOTAL NO. OF MARKS
Art	29	32	29	8	2	302
Commercial	21	39	33	3	4	348
English	10	27	28	22	13	984
History	13	27	36	16	8	710
Home Economics	21	36	28	10	5	288
Languages	19	32	24	7	18	323
Mathematics	9	27	32	22	10	596
Mechanical Arts	26	45	20	6	3	462
Music	40	36	16	8	0	684
Physical Education	18	58	20	3	1	886
Science	22	33	30	9	6	469
All Departments	20	36	26	11	7	6016
Junior High School	15	33	29	16	7	2966
Senior High School	26	31	26	9	8	2153
Academic Departments	13	28	31	17	11	3046
Non-academic Departments	26	44	22	6	2	2970

[1]Carl E. Hendrickson: *School Marks at Van Nuys High School, Educational Research Bulletin,* Los Angeles City Schools, Vol. VII, No. 4 (December, 1927), pp. 8-9.

Hendrickson's comment on the foregoing table is both concise and adequate to the facts when he says:

> It is probably significant to point out that 56 per cent of the total marks were A or B or college recommending and only 18 per cent were D and E. In other words the distribution is skewed over toward the highest marks considerably. Here it may be of interest to add that all the mental and educational tests given in this school result in approximations to the normal curve. Also the level of intelligence here is about normal.

The author's study of the marks in a small high school. Table 5 presents a study of 659 school marks for a six-week period in the University of Oregon High School, Eugene, Oregon. P. T. refers to the grades of all practice teachers grouped together. The totals include regular (designated A, B, C, etc.) teachers and practice teachers. The school standards were adopted in faculty meeting as the official standards of the school.[1]

TABLE 5

LETTER GRADES ASSIGNED BY TEACHERS IN THE UNIVERSITY HIGH SCHOOL, EUGENE, OREGON, FOR A SIX-WEEK REPORT PERIOD

TEACHER	PERCENTAGES				
	A	B	C	D	E
A	48.7	44.7	6.6	0.0	0.0
B	28.6	58.6	12.8	0.0	0.0
C	15.2	66.3	15.7	1.7	1.1
D	41.8	48.6	6.7	2.9	0.0
E	8.9	78.6	12.5	0.0	0.0
F	30.0	55.0	15.0	0.0	0.0
P. T.	46.8	37.0	8.4	3.2	4.5
Total	32.0	54.0	11.0	1.6	1.4
School Standards	6.25	25.0	37.5	25.0	6.25

"The grades as a whole are badly skewed upwards despite the existence of a school standard defining marks thus:

[1]Quoted from *The Improvement of the Written Examination*, pp. 47-48.

A to represent the upper 6% of the pupils, approximately,
B to represent the next 25% of the pupils, approximately,
C to represent the middle 35%-40% of the pupils, approximately,
D to represent the next 25% of the pupils, approximately,
E to represent the lowest 6% of the pupils, approximately.

"Teacher E illustrates an interesting situation. At the previous report period, each teacher's grade distribution together with a graph had been posted in the faculty room. This teacher's summary showed more than fifty per cent of A grades given. In the effort to overcome this situation, this teacher forced down the number of A's to a point well below every other teacher in the school, apparently by the very simple expedient of changing the A's to B's, a solution which did not help matters greatly for the distribution taken in its entirety! Could one take these grade distributions at face value, he could not but be impressed with the truly wonderful efficiency of a school where more than eighty-five per cent of the pupils earned either A's or B's. Kelly[1] (after extensive study of prevailing marking systems) has summarized his conclusions in these words:

A given grade or mark means many widely different things to different teachers when they are rating pupils for promotion. As measured by the achievement of the several school groups in their later work this difference amounts in some cases to as much as the difference between a G (good) and F— (fair minus) in elementary schools where the basis of marking includes only the steps P, F, G, and E. In high schools there is enough difference between the standards of schools as wholes that, measured by the achievement of the school groups in later school work, a mark of 70 in one school means more than a mark of 81 in another school having the same passing standard by points. Within the high school and within the college the percentage of pupils which the various instructors fail as a common practice extending over several years varies from 0 to 28, or more.

Comment on the three investigations of teachers' marks. The three studies presented show the same general tendencies. No significance is to be attached to variations in the absolute numbers of different letter grades given in the three

[1] *Op. cit.*, page 133.

schools since the bases for distributing marks naturally differ as a matter of administrative policy. Within a given school, however, there should be a reasonable equality of percentages of letter-grades from one department to the next.

First let us consider what an "A" means. *The number of A's which should be given is wholly a matter of definition.* Each school must settle for itself such questions as a matter of administrative policy. There is an idea current that such letter-grade distributions are somehow derived from and fixed by the normal curve. The normal curve is totally impotent in the matter, unless we except the fact that the normal curve does suggest the relative proportions of letter-grades in contrast with the absolute numbers. To illustrate:

	A	B	C	D	E
Case I	5	20	50	20	5
Case II	10	25	50	10	5

Case I is obviously more in harmony with the observed facts about the distribution of individual differences and the phenomena of organic variation generally. Case II violates these facts due to the marked skewness or lack of symmetry in its distribution. To this extent the normal curve is a rough guide. But, if the question is which *case* gives the proper percentages of A's (or any other letter-grade), the normal curve helps not at all.

A uniform marking system for the entire United States would have far-reaching advantages, but such uniformity is little more than a "pious wish." Each school will probably continue, for reasons of greater or less weight, to distribute marks according to its locally adopted scheme. *The important thing is to adopt some arbitrary standard* (it must be arbitrary), *and then adhere to it, department by department and teacher by teacher, with the one important qualification that such standards are just only in the long run.* The validity of

such marking plans rests upon large numbers. It must not be applied mechanically to small classes. But small classes become large classes when the results are pooled semester after semester and year after year! Considerable departures from the adopted distribution must be permitted to the teacher (provided she can prove her point by reliable objective evidence) in marking individual classes, but such latitude must not be allowed to operate *systematically* time after time. *It is possible, but not probable,* that the Latin teacher might really have had but three per cent of A-pupils over a ten-year period, while the civics teacher had seventeen per cent in the same time. Such a finding is possible, but not *typical.* There are more probable explanations.

If four conditions were approximated closely, grades might be assigned by the mechanical application of some adopted percentage-letter-grade plan. These conditions are:

1. That very large (theoretically, infinite) numbers of pupils are to be graded.
2. That there is no selective enrollment in different classes, sections, etc. (This implies chance assignment to classes, absence of ability-grouping, and non-existence of selection upon a basis of ability in electing programs of study.)
3. That teachers are equally efficient in their instruction.
4. That marks are based upon wholly valid and reliable measurements.

The problem is to decide whether such wide divergencies as were shown in the studies of Johnson, Hendrickson, and Ruch are fully explained by the factors of small populations, selective enrollment, differences in efficiency of instruction, etc., or whether it is more reasonable to suppose that much of the departmental variation is to be explained by such facts as non-adherence to the accepted school standards, differences in the subjective standards of different teachers, etc.

The reader can adjudge the relative merits of these rather intricate differences in points of view. When, as the author found in a study not reported in detail here, one department in a large university gave fifteen per cent of A's and another gave .9 of one per cent of A's, a problem exists. In this particular case the former department gave as its alibi the statement that superior students were attracted to its courses. The intelligence test records of the university were consulted. The average of the students of this department was well below the average of the university! Taking the average marks and the average intelligence ratings by departments, only the slightest correlation was found. This leads us to suspect that variations in the abilities of pupils from one department to the next are probably much too small to explain the variations in marks.

We have a certain type of teacher who has "high standards." She prides herself on having her "sights" high. In extreme cases she boasts that she has never given an A. A correspondent once wrote the author about a college student who grew tired of receiving a constant string of C's. He put the teacher to an (underhanded and ungentlemanly?) test by copying word for word an exquisite bit of literature by a prominent author. This he handed in as his own work. He received a C! Such a teacher may have had high standards; who can say? Again, it may merely have been chronic indigestion or ignorance.

Adherence to an adopted grading system is nothing more than "playing the game fairly." No one holds such grading plans to be more than definitive. Uniformity of practices has its obvious advantages. Departures from uniformity must be left open, with the qualification that the burden of proof is on the one who departs. The essential thing is assurance that the underlying evaluations are valid and reliable; that the final grades are based upon what has been termed a defensible rank-order of abilities. If marks are relatively fair, the exact final distribution reduces to a pure

matter of definition of school policy. Tests and examination, if well-constructed, form a good basis for both adherence to and departure from the defined standards. Personal judgments also have their values, but tests can be made to discriminate finer differences than can unaided subjective estimates.

INVESTIGATIONS OF REGRADINGS OF THE SAME PAPERS

The studies of Starch and Elliott. The pioneer work in this field was done by Starch and Elliott,[1] who submitted exact copies of the same examination papers to a large number of teachers. Figure 5 shows the marks of 142 English teachers on the same examination in English; Figure 6 shows the variations observed when 115 teachers graded the same paper in geometry.

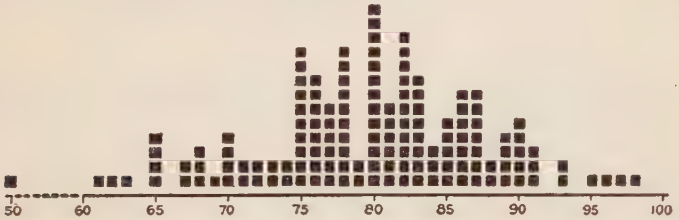


FIG. 5.—The marks assigned to the same English paper by 142 teachers of English (after Starch).

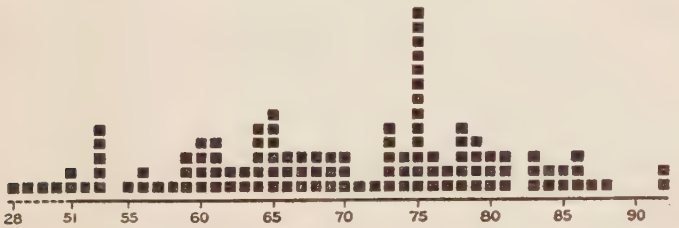


FIG. 6.—The marks assigned to the same paper in geometry by 115 teachers of high-school mathematics (after Starch).

[1] *School Review*, Vol. XX: pp. 442-57; Vol. XXI: pp. 254-59; Vol. XXVI: pp. 676-81.

It had previously been supposed that marks in mathematics could be much more objectively given than for such subjects as English and history. Rather strangely it turned out that the English papers were graded with somewhat greater uniformity than those in history and mathematics in this investigation. Later, Starch found that teachers within the same school show almost as great differences in their markings as do teachers selected from different schools.

Ruch's repetition of the studies of Starch and Elliott.[1] "The first experiment was that of submitting three answers to the same question in geography to about a hundred teachers for regrading. All three answers were taken from the papers of one class, those papers selected being the best, poorest, and median papers. The pupils' responses were copied verbatim and mimeographed, retaining all errors. The instructions to the teachers and the answers are reproduced below.

Directions: Below are three actual answers to the question: *Name and locate five of the largest cities of the United States and name their leading industries, exports, and imports.*

Grade each of the three answers on a scale of 0 to 20, according to your best judgment of its merit, 20 being an answer ordinarily accepted by teachers as entirely satisfactory, and 0 being an answer practically without discernible merit.

ANSWER 1

Five of the largest cities in the United States is Detroite. An export is Cars. And industry is Manufactoring. Chicago is an important city and an export is Manufactored and canned goods. An industry of Chicago is meate packing. New York is another important city. An industry of N. Y. is manufactoring. An export of N. Y. is manufactored goods. Pittsburge is an important city of U. S. An export is iron ore. A industry of Pittsburge is manufactoring. Another important city of U. S. is New Orleans. An export of New Orleans is cotton. An industry of New Orleans is manufactoring.

Grade.....

[1]Quoted with slight changes from *The Improvement of the Written Examination*, pp. 55-60.

ANSWER 2

The five largest cities of the United States are (1) navada (2) Arkansas. The leading industry of navada is manufacture and the leading industry of Arkansas is agriculture. The leading imports are manufacturing mostly.

Grade.....

ANSWER 3

The 5 largest cities of United States are New York, Chicago, St. Louis Boston, San Francisco. New York is in the State of New York. Chicago is in the State of Illinois. St. Louis in in the state of Missourri. Bostin is in the state of Mass. San Fransisco is in the state of California. New York is a mnufacturing city. Chicago is noted for meat packing center. St. Louis is noted for manufacturing textile goods and iron goods. Boston is noted for manufacturing of textile and iron goods. San Fransisco is noted for the packing of fruit. New York exports iron goods and imports wool, cotton, and other raw materials. Chicago exports meat and hides and grain and imports food and grains. St. Louis exports manufactured products and imports raw materials.

Grade.....

"Table 6 on the next page summarizes the regradings of the three answers by ninety-one teachers. The facts illustrate several points previously brought out. In the first place it is to be noted that the original grading in no case agrees at all closely with the average grades of the ninety-one teachers. The difference is from two to five points in each case. The ninety-one teachers show a wide variance of opinion about the merits of these three answers. Which set of marks is correct, the originals or the regradings?

"The only answer is to accept the average of the group as approximating the truth. Granting this for the moment, we find that the original mark of Answer 1 was about five points too high, and that for Answer 3 was about four points too low. What reason can be assigned for this situation? So far as the facts presented in Table 6 go, it is impossible to answer this question satisfactorily. A probable explanation is to be found in the fact that Answer 1 was taken from a

rather superior paper, but Answer 3 was found in a decidedly inferior paper, the papers taken as a whole. The original grader has undoubtedly been unconsciously influenced by these facts; thus tending to grade leniently a poor answer in an otherwise very good paper, and underestimating the value

TABLE 6

ORIGINAL MARKS AND MARKS ASSIGNED BY 91 TEACHERS WHO REGRADED THREE ANSWERS TO A QUESTION IN GEOGRAPHY

MARK	ANSWER 1	ANSWER 2	ANSWER 3
20	1	..	9
19	0	..	3
18	1	..	21
17	1	..	12
16	2	..	17
15	10	..	15
14	2	..	3
13	8	..	1
12	14	..	3
11	5	..	1
10	24	..	3
9	5	..	1
8	7	..	1
7	3	..	0
6	4	..	0
5	1	..	0
4	0	..	0
3	2	..	1
2	1	3	0
1	0	1	0
0	0	87	0
Number	91	91	91
Mean	10.9	0.1	16.1
Standard Deviation	3.2	0.4	2.9
Original Mark	16	2	12

of a good answer which formed a part of an inferior paper. This same fact was evident many times in these studies of examination papers, and if space permitted, considerable statistical proof of the operation of such unconscious biases could be presented. Similarly, the existence of systematic

tendencies for certain teachers to grade too leniently and for others to be too harsh in their judgments might be amply demonstrated by facts at hand. These phenomena are too well recognized by teachers, however, to require proof or comment.

"In a second experiment, three entire history papers taken from a seventh-grade class in American history were mimeographed so as to preserve all spelling and grammatical errors, and as much of the mechanical form as was possible in the duplication. The papers were again the best, median, and poorest papers of the class. These were then regraded by 115 teachers, independently. The original marks were, of course, not known to the group. Space does not permit the recording of either the questions or the copies submitted to the teachers. Table 7 on the next page summarizes the facts of this experiment, which was similar in nature to those formerly reported by Starch and Elliott.

"Table 7 would seem to show that the original grader of these papers was too tolerant in her standards, the composite judgments lowering all of the original marks by at least ten points. This illustrates the operation of systematic biases which have been asserted as a source of error. Marked variability exists as before in the judgments of the 115 teachers as to the worth of these papers. In this case, however, the ranks of the papers remain the same as before, so that there is little doubt that distinguishable differences in merit are present among the three examinations. It should be recalled that these three papers are as widely spaced among the class as was possible, the best, the median, and the poorest paper being selected. Attention should also be called to the amount of overlapping present, a fact that is more graphically shown in Figure 7 on page 83. Only the average of a very large number of teachers' markings would demonstrate decisively that real differences in merit are present in these papers.

TABLE 7

ORIGINAL MARKS AND MARKS ASSIGNED BY 115 TEACHERS WHO REGRADED THREE PAPERS FROM A CLASS IN AMERICAN HISTORY

MARK	PAPER 1	PAPER 2	PAPER 3
100	6	..	..
95	33	..	..
90	32	1	..
85	22	12	1
80	15	13	4
75	6	29	9
70	1	18	5
65	..	16	20
60	..	12	16
55	..	9	19
50	..	3	13
45	..	2	8
40	..	..	14
35	..	..	5
30	..	..	0
25	..	..	1
Number	115	115	115
Mean	88.7	70.3	56.6
Standard Deviation	6.6	9.9	12.3
Original Mark	100	88	67

Wood's studies of the examinations of the College Entrance Examination Board. The following paragraphs are quoted directly from Wood.[1]

"More recent evidence of the inadequacies of the marks derived by the traditional examinations was brought out in a study of the algebra and geometry examinations of the College Entrance Examination Board for June, 1921. About four hundred algebra and an equal number of geometry papers selected at random were scored each twice independently by two different readers of the Board. The correlations between the first and second scorings were very high for both algebra and geometry, about .98 and .96.

[1]B. D. Wood, *Measurement in Higher Education* (Yonkers-on-Hudson, New York: World Book Company, 1923), pp. 124-125.

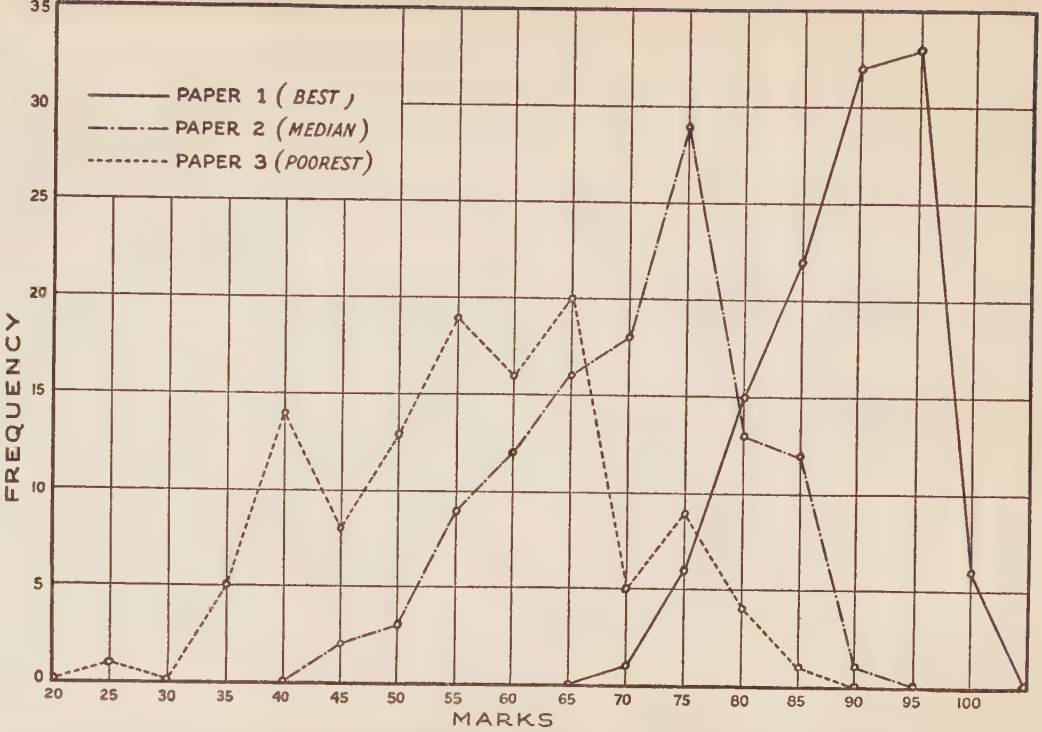


FIG. 7.—Curves showing the distributions of the marks of 115 teachers who regraded the three history papers in Table 7.

In spite of this almost perfect objectivity in the scoring, however, the reliabilities of the examinations themselves were very low. The correlation between random halves of the algebra examination was found to be only .61, and that between random halves of the geometry examination only .41. By the use of Brown's formula, the reliability of the whole algebra examination is estimated as not greater than .76, and that of the geometry examination as not greater than .58.

"The meaning of these reliability coefficients may be made clearer by a consideration of a hypothetical case closely resembling the actual situation faced by the Board in giving college entrance examinations.

"Let us suppose that ten thousand candidates are tested with Form A of a given geometry examination whose reliability is about .60, and that thirty per cent of the ten thousand fail. Now let us suppose that the same ten thousand are tested with another equivalent geometry examination, say Form B, whose reliability is also .60, and which fails thirty per cent of the candidates.

"If the reliability of the two forms of the examination were 1.00, the same three thousand candidates would be failed by both forms; but with a reliability of only .60, the agreement on failures would be as follows in gross numbers:

Failed Form A Passed Form B 1279	Passed on both Forms 5721
Failed on both Forms 1721	Passed Form A Failed Form B 1279

"In other words, accepting the results of one such examination as valid, which was done by the Board in 1921, another equivalent examination would pass 1279 of the 3000 failed by the first, and would fail 1279 of the 7000 passed on into college by the first.

"If we assume that fifty per cent were failed by each of the two forms, the displacements in gross numbers would be:

1015	3985
3985	1015

The fate of two thousand in ten thousand candidates would be reversed by an equivalent form of the same examination when the reliability is no higher than that of the C.E.E.B. Mathematics C Examination for June, 1921."

Another view. Recently Bolton has boldly denied that teachers show marked lack of uniformity in marking papers.[1] His evidence is based upon an investigation which he conducted together with a re-examination of a minor study of Starch (not the larger ones previously reported in this volume). Bolton's experiment was very well conceived except in one or two respects noted below. His statistical arguments are not nearly so fortunate.

Bolton had a number of sixth-grade arithmetic teachers make examinations and administer them. From the papers he selected by a sampling process the results for twenty-four pupils. These papers were then graded by twenty-two teachers. This procedure is admittedly sound save in the very important respect *that Bolton selected what is probably the second most highly objective school subject (arithmetic) for his investigation.* Starch, it will be recalled, chose highly subjective school subjects in the main. So did most of those who repeated the work of Starch. Bolton's point of view which guided his set-up for the investigation can best be expressed in his own words. Speaking of the type of teachers used by previous investigators, he says (p. 24):

> They vary in experience; their everyday work may vary from teaching beginners to read to administering a school system with a hundred teachers;

[1] F. E. Bolton, "Do Teachers' Marks Vary as Much as is Supposed?" *Education*, Vol. XLVIII (1927), pp. 23-38.

some teach one subject, some many others; some have had real professional training, some absolutely none. Possibly not *one tenth* [italics mine] of those marking the papers have had experience in marking papers in that subject, and many are so rusty in the facts of that subject as not to know the answers to the questions themselves.

While it is possible that Bolton's statements may be true in some cases, the assertion that nine-tenths of the teachers taking part in the experiments of Starch (and the present author) were incompetent is gross misrepresentation. Starch used 142 English and 115 geometry teachers from the North Central Association selected under instructions to have the grading done by the "principal teacher" of the subject. The total of 206 teachers used by the author in grading the geography and history replies of Tables 6 and 7 were all selected upon the basis of experience and training, and all had more than average professional training.

The main objection to Bolton's procedure and his conclusions rests in his choice of statistical methods. He averaged the marks of the twenty-two teachers for each of the twenty-four papers, obtaining the values in column (*e*) of Table 8. He next found the average of the deviations about such averages, column (*f*). The other columns are the selections of the present author from Bolton's Table I.

There is of course no objection to the use of averages and average deviations from such averages as a statistical procedure unless we question the *choice* of such a method of interpretation. After all, the range between the highest and lowest marks given an individual paper may be the important thing, and not the fact that the average deviations about the averages of twenty-two teachers is fairly small.

Now it must be admitted that Pupil 12 was fairly marked, for all practical purposes, by any one of the twenty-two teachers. In fact, he is the only one in the whole lot about whom such an assertion may be made unqualifiedly. But how about Pupils 5, 6, 13, 17, 19, 21, and 23, not to mention

TABLE 8

SELECTED DATA FROM BOLTON'S STUDY OF TEACHERS' MARKS

(a)	(b)	(c)	(d)	(e)	(f)
PUPIL	LOWEST MARK ASSIGNED	HIGHEST MARK ASSIGNED	RANGE	AVERAGE OF 22 TEACHERS	AVERAGE OF DEVIATIONS
1	77.5	100	22.5	88.7	3.6
2	75	90	15	85.0	4.3
3	70	95	25	88.7	3.5
4	43	71	28	57.8	5.9
5	65	95	30	84.6	6.4
6	25	70	45	51.0	10.5
7	74	91	17	84.8	3.1
8	73	93	20	84.7	5.4
9	85	100	15	93.8	2.7
10	71	95	24	89.6	3.4
11	66	80	14	77.5	2.8
12	84	90	6	88.0	1.4
13	46	85	39	71.5	7.8
14	67	84	17	74.4	2.3
15	47	74	27	62.7	4.6
16	82	98	16	91.5	3.9
17	53	95	42	78.9	8.4
18	70	91	21	77.9	4.4
19	37.5	78	40.5	53.5	7.2
20	43	71	28	55.3	4.0
21	48	88	40	74.5	8.5
22	35	58	23	45.1	4.7
23	43	75	32	65.3	7.3
24	17.5	47	29.5	27.2	5.8
Medians	65.5	89	24.5	77.7	4.5

certain others? If Pupil 6 was graded all the way from 25 to 70 with an average of 51, just what are we to conclude about his ability? The median of the ranges is roughly twenty-five points. In about half the cases the most lenient teacher marked from twenty-five to forty-five points higher than the most severe one.

Do such data support the statement, "A glance at the distribution . . . of variations from the average *discredits entirely* [italics mine] the assertion that there is no uniformity of marks" (p. 28)?

Starch, for much more subjective subjects than arithmetic, reported ranges among the marks of more than one hundred teachers as follows:

Subject	Range	No. of Teachers
English	50 to 98 (48 points)	142
Geometry	28 to 92 (64 points)	115

The author found results for three history papers as follows:

Paper	Range	No. of Teachers
1	70 to 100 (30 points)	115
2	45 to 90 (45 points)	115
3	25 to 85 (60 points)	115

Taking all facts into consideration, the following statements are given with the hope that the reader will evaluate each and arrive at some decision as to the validity of Bolton's refutation of Starch and others:

1. Bolton dealt with a reasonably objective school subject; spelling being, perhaps, the only less subjective elementary school ability.

2. He used a much smaller number of teachers (22), thus possibly reducing the variability considerably.

3. He shifted the argument from the idea of extreme differences (ranges) to deviations about an average. This is defensible, of course, speaking purely statistically, but the fact remains that his interpretation is not comparable to that of Starch. When comparable treatments are made, the differences between Bolton and Starch are not so very great. This leaves us squarely with the question whether the important thing is what the average teacher does or what the extremes of individual teachers do. Take Pupil 10 of

Table 8, who represents about the average situation found by Bolton. One teacher gave him 95, another 71. The former grade would probably be an A and the latter a D or an E on a five-point scale. The blunt fact remains that it makes a lot of difference to that pupil which school he attended and what teacher he "drew." With the admitted exception of Pupil 12, and possibly (to be generous) Pupils 2, 7, 9, 11, 14, and 16, the other seventeen have a right to "kick" about the situation. And, Pupils 5, 6, 13, 17, 19, 21, and 23 have the moral right to riot and insurrection. They are failures or successes depending upon their teachers, regardless of averages and average deviations.

The reader must judge for himself whether or not Bolton has made his point.

STUDIES ON THE RELIABILITY COEFFICIENTS OF EXAMINATIONS

Introduction. Reliability has already been discussed and defined in several ways (Chapter II). The quantitative expression of the degree of reliability is usually made in terms of coefficients of correlation. Correlation is, as the term itself suggests, co-relation or the degree of correspondence between two series of numerical values. The mathematics of the computation of coefficients of correlation will be reserved for Chapter XV, which discusses the elementary statistical methods related to examination practices. For the present it will serve our purposes to know the general significance of the "reliability coefficient."

When two sets of measures of the same ability or function are correlated, we term the resulting coefficient of correlation a reliability coefficient. By "two sets of measures of the same ability or function" we have in mind equivalent or comparable forms of the same test, or some closely analogous pair of measures.

If a teacher gives two forms of a standard test or if she administers two duplicate examinations, the two sets of scores may be compared by correlation, the resulting coefficient in this case being a *reliability coefficient.*

There are several ways of obtaining reliability coefficients when we are studying examinations:

1. Two equivalent (or roughly equivalent) tests may be given and the results correlated.

2. A single test may be given, the papers graded independently by two teachers, and the two sets of marks are then correlated.

3. A single examination, graded by a single person, may be broken into two half-examinations by some chance method (e. g., taking alternate items in each half-form), and the halves are then correlated. (This gives directly the reliability of half the examination. The reliability of the whole examination can be estimated rather accurately by the use of the appropriate mathematical formula.)[1]

These three methods are not exactly comparable in meaning, but each has its distinct uses in attacking the question of the reliability of examinations.

"High" and "low" reliability. When is a correlation "high" or "low"? There are as many answers as there are textbooks on statistics and measurement. The question is far too intricate for full discussion. Instead, we shall beg the issue by laying down dogmatic statements which will define the author's point of view for purposes of present interpretations.

Correlations of 0.00-0.25 are insignificant.
Correlations of 0.25-0.50 are low.
Correlations of 0.50-0.80 are fairly significant.
Correlations of 0.80-0.95 are fairly high.
Correlations of 0.95-1.00 are high.

[1]See Chapter XV for a discussion of the use of the Spearman-Brown formula in this connection.

TABLE 9

SUMMARY DISTRIBUTION OF COEFFICIENTS OF RELIABILITY FOR WRITTEN EXAMINATIONS

SIZE OF COEFFICIENT OF CORRELATION	FREQUENCY
.95	1
.90	2
.85	4
.80	4
.75	9
.70	4
.65	9
.60	8
.55	4
.50	4
.45	5
.40	2
.35	1
.30	4
.25	1
.20	0
.15	1
.10	0
.05	1
.00	0
−.05	0
−.10	0
−.15	1
−.20	1
Total	66
Median .65	

These statements are something of a compromise between strict statistical considerations and the more practical question of the frequency with which tests and measurements attain these several levels of magnitude. In general, the present interpretation is more conservative than that found in most textbooks.

Monroe and Souders's study of the reliability of written examinations (traditional type). Monroe and Souders computed reliability coefficients for sixty-six examinations. The same pupils were given two examinations; "in most cases the questions were prepared and the papers marked by different teachers." Table 9 is reproduced from Monroe.[1] The range of reliability coefficients was from −0.20 to 0.95; the median being 0.65.[2]

McGregor and Ruch's study of state eighth-grade examinations.[3] "Requests were sent to all state superintendents of public instruction for copies of all official state eighth-grade examinations for as many years past as possible.

[1]W. S. Monroe, J. C. DeVoss and F. J. Kelly, *Educational Tests and Measurements* (Revised edition, Boston: Houghton Mifflin Co., 1924), p. 471.

[2]Negative correlations are interpreted in exactly the same ways as positive ones except that the relationships show tendencies to be inverse; i. e., the pupils doing well on one examination *tended* to do poorly on the other; at least, the small amount of correlation noted in such negative situations was of this inverse sort.

[3]Quoted with minor changes from Ruch, *et al: Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Co., 1926), pp. 6-12.

Eleven states responded with questions which were actually used in this investigation. A few other states delayed their returns until too late for inclusion.

"The examination questions from the eleven states were classified in three groups: United States history, geography, and civics (citizenship). Key numbers were assigned to the examinations in order that the source of the questions would not be revealed to the schools co-operating in the experiment. These key numbers are used in the tabulations to be presented in this chapter. Every attempt was made to avoid any publicity about the particular states furnishing the questions, since it was the intention of the invesitgators to study the eighth-grade examination system as a whole rather than to direct attention to the examination practices of particular states. Parenthetically, it may be said that no evidence was found which suggested that any of the individual states were measurably superior or inferior to the others in the character of the examinations set for their eighth-grade pupils.

"Occasional questions were omitted when such questions were based upon local history, geography, or government, for two reasons: (1) in order not to reveal the source of the examination, and (2) because the inclusion of such questions would not be a valid procedure in view of the fact that the questions would be used in states other than the one for which the examination was devised. Since the examinations usually offered some degree of choice in the questions to be answered, it was possible to omit an occasional question without much violence to the examination.

"The sets of questions were then mimeographed so that each pupil might have his individual copy. The following directions were given to the pupils:

You are to be given an examination in ..... (the teacher supplied the subject), which pupils in another state had to take

in order to get their eighth-grade diplomas. We want to see if you can do as well as pupils in other states. Work as fast as you can without making mistakes. When you have finished, record your time in a square which you should make on the last page near the bottom."

"The teacher timed the examinations to the nearest one-half minute by means of the plan of writing the elapsed time at half-minute intervals on the blackboard. All of the pupils used in the experiment wrote on two examinations for the same subject, viz., the set of questions for the year 1923 and the set for 1924. Thirty-two experienced teachers of the social studies did the scoring, *every paper being marked independently by two teachers.*

"That the investigation included a wide sampling of state examinations, pupils, and scorers is shown by the following facts:

(1) The eighth-grade examinations were drawn from eleven different states.

(2) Thirty-two different sets of questions were used, i. e., both the 1923 and 1924 questions for sixteen school subjects.

(3) Thirty-two different teachers read the papers, each teacher reading the 1923 and 1924 examinations for one class of pupils.

(4) The papers include two examinations each from 952 pupils representing 15 schools and 11 states.

"All papers were graded upon a basis of 100%. If the examination included ten questions, each question was allowed a maximum of ten points. Where five, eight, etc., questions were employed, the 100 points were divided evenly among the questions.

"Treatment of the results. The sixteen examinations permitted the calculation of a total number of ninety-six correlations, each correlation being a *reliability coefficient* from some point of view. The six correlations possible for each set of examinations may be shown by the following outline:

(1) 1923 examination: scorer No. 1 vs. scorer No. 2.
(2) 1924 examination: scorer No. 1 vs. scorer No. 2.
(3) Scorer No. 1: 1923 examination vs. 1924 examination.
(4) Scorer No. 2: 1923 examination vs. 1924 examination.
(5) 1923 examination scored by No. 1 vs. 1924 examination scored by No. 2.
(6) 1924 examination scored by No. 1 vs. 1923 examination scored by No. 2.

"The six numbered columns of Table 10 correspond to the above numbering scheme. Table 10 presents the ninety-six reliability coefficients possible for the sixteen sets of examinations.

TABLE 10*

RELIABILITY COEFFICIENTS OF 16 STATE DIPLOMA EXAMINATIONS, YEAR (1923) AGAINST YEAR (1924) AND SCORER AGAINST SCORER

No.	KEY	SUBJECT	(1)	(2)	(3)	(4)	(5)	(6)	POP.
1	G-2	Ele. Citizenship	.45	.21	−.05	.46	.34	−.26	102
2	I-1	U. S. History	.60	.43	.16	.41	.23	.17	31
3	J-1	U. S. History	.47	.30	.44	.73	.25	.22	32
4	F-3	Geography	.58	.39	.37	.22	.17	.55	36
5	F-1	U. S. History	.89	.99	.67	.64	.67	.69	94
6	D-2	Civics	.82	.88	.22	.25	.33	.23	36
7	I-3	Geography	.40	.88	.32	.48	.29	.41	32
8	M-2	Civics	.80	.82	.47	.55	.46	.52	61
9	D-1	U. S. History	.81	.22	.54	.65	.49	.35	34
10	L-1	U. S. History	.79	.57	.73	.48	.45	.66	107
11	A-1	U. S. History	.81	.85	.66	.71	.56	.65	42
12	B-1	U. S. History	.53	.58	.36	.34	.27	.41	97
13	B-2	Civics	.63	.20	.36	−.18	−.06	.25	82
14	K-1	U. S. History	.93	.91	.56	.67	.68	.51	99
15	I-2	Civics	.81	.53	.37	.27	.52	.46	35
16	E-1	U. S. History	.75	.12	.26	.59	.60	.19	32
Averages			.69	.56	.40	.45	.39	.38	(952)
Averages by Pairs of Columns			.62		.43		.38		

*COLUMN (1): 1923 examination, scorer No. 1 vs. scorer No. 2.
COLUMN (2): 1924 examination, scorer No. 1 vs. scorer No. 2.
COLUMN (3): Scorer No. 1, 1923 examination vs. 1924 examination.
COLUMN (4): Scorer No. 2, 1923 examination vs. 1924 examination.
COLUMN (5): 1923 examination scored by No. 1 vs. the 1924 examination scored by No. 2.
COLUMN (6): 1924 examination scored by No. 1 vs. the 1923 examination scored by No. 2.

"Table 11 shows the average scores or marks assigned to both 1923 and 1924 examinations of the sixteen state diploma examinations.

TABLE 11

AVERAGE SCORES (MARKS) ASSIGNED BY TWO DIFFERENT SCORERS FOR BOTH THE 1923 AND 1924 EXAMINATIONS (THE 16 STATE DIPLOMA EXAMINATIONS)

No.	Key No.	(1) 1923 Exam. Scorer 1	(2) 1923 Exam. Scorer 2	(3) 1924 Exam. Scorer 1	(4) 1924 Exam. Scorer 2
1	G-2	67.5	82.0	73.0	70.4
2	I-1	71.7	67.1	57.0	71.0
3	J-1	68.1	43.6	64.9	45.6
4	F-3	70.0	54.8	70.3	69.9
5	F-1	47.7	45.3	51.4	41.9
6	D-2	55.7	47.5	65.6	59.1
7	I-3	51.0	62.3	50.4	68.9
8	M-2	51.0	48.6	48.5	42.9
9	D-1	38.3	34.4	48.5	30.7
10	L-1	49.3	56.5	38.3	65.3
11	A-1	42.5	28.1	25.3	18.6
12	B-1	14.4	7.7	24.4	25.3
13	B-2	29.3	26.0	24.8	11.5
14	K-1	48.1	59.0	61.3	64.7
15	I-2	38.3	41.4	68.1	58.0
16	E-1	21.0	26.9	8.6	12.4

SUMMARY OF DIFFERENCES*

	(1–2)	(3–4)	(1–3)	(2–4)	(1–4)	(2–3)
Average Difference....	8.6	9.4	9.2	9.9	11.4	12.0
Largest Difference....	24.5	17.6	29.8	27.0	23.9	26.8
Smallest Difference...	2.4	2.0	0.3	0.4	0.1	0.2

"Table 12 shows the differences in the average scores (of Table 11 assigned by two different scorers for both forms of the 16 state diploma examinations. Algebraic signs are ignored. The outline preceding Table 12 is necessary in interpreting the meanings of the columns lettered (*a*), (*b*), (*c*), etc., in Table 12."

*(1–2), (3–4), etc., refer to the differences in the columns numbered 1, 2, 3, and 4.

(*a*) Differences in the average scores assigned to the 1923 and 1924 examinations by scorer No. 1.
(*b*) Differences in the average scores assigned to the 1923 and 1924 examinations by scorer No. 2.
(*c*) Differences in the average scores assigned to the 1923 examinations by scorers Nos. 1 and 2.
(*d*) Differences in the average scores assigned to the 1924 examination by scorers Nos. 1 and 2.
(*e*) Differences in the average scores assigned when scorer No. 1 read the 1923 examination and scorer No. 2 read the 1924 examination.
(*f*) Differences in the average scores assigned when scorer No. 1 read the 1924 examination and scorer No. 2 read the 1923 examination.

TABLE 12

DIFFERENCES IN THE AVERAGE SCORES (OF TABLE 11) ASSIGNED BY TWO DIFFERENT SCORERS FOR BOTH FORMS OF THE 16 STATE DIPLOMA EXAMINATIONS Algebraic Signs are Ignored.*

No.	KEY NO.	(*a*)†	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)
1	G-2	5.5	11.6	14.6	2.5	2.9	9.1
2	I-1	14.7	3.9	4.6	14.0	0.7	10.1
3	J-1	3.2	2.0	24.5	19.3	22.5	21.4
4	F-3	0.3	15.1	15.2	0.4	0.1	15.5
5	F-1	3.6	3.3	2.5	9.4	5.8	6.1
6	D-2	9.9	11.6	8.2	6.5	3.4	18.1
7	I-3	0.6	6.6	11.3	18.5	17.9	11.9
8	M-2	2.5	5.7	2.4	5.6	8.1	0.2
9	D-1	10.1	3.7	4.0	17.8	7.6	14.1
10	L-1	11.0	8.8	7.2	27.0	16.0	18.1
11	A-1	16.7	9.4	14.4	7.2	23.9	2.3
12	B-1	10.1	17.7	6.7	0.9	10.9	16.8
13	B-2	4.5	14.5	3.3	13.3	17.8	1.2
14	K-1	13.2	5.8	10.8	3.4	16.6	2.4
15	I-2	29.8	16.6	3.0	10.1	19.7	26.8
16	E-1	12.3	14.5	5.9	3.7	8.6	18.2
Averages		9.2	9.4	8.6	9.9	11.4	12.0
Averages by pairs of columns		9.3		9.2		11.7	

*See Table 10 for subjects involved and numbers of cases.
†See comments above the table for description of the columns lettered, (*a*), (*b*), (*c*), etc.

Gordon's study of the New York Regents' Examinations.[1] Gordon, working under the direction of the author, carried on an investigation similar to that reported by McGregor and Ruch except that examinations prepared by the New York Regents were employed. Table 13 shows certain of the findings.

It should be noted that the studies of McGregor and Ruch and Gordon used examinations constructed for use in one state but which were applied to pupils in other states. This resulted in lowered average scores, and very possibly in somewhat reduced reliability coefficients. This irregularity can hardly be held to invalidate the results completely. It was further true that the regularly appointed official readers were not used in these two investigations.[2]

Wood's studies on old-type test reliabilities. Dr. Ben D. Wood has been a most indefatigable investigator of the comparative validities and reliabilities of both old- and new-type examinations. His studies have taken many different directions, and unfortunately he has not found time to summarize his findings in any one reference work. A series of investigations has attacked in turn the examinations of Columbia University, those of the New York Regents, those of the College Entrance Examination Board, and certain unofficial and individual examinations. On the points at issue in this chapter, Wood is perhaps the outstanding authority; certainly he has been the most prolific and consistent worker. The author is taking the liberty of quoting occasional statements from the work of this investigator rather than summarizing fully any one study.

[1]W. E. Gordon, *A Study of the Reliability of Examinations Based upon the New York Regents' Examinations in the Social Studies*, Ph. D. Dissertation (1925), State Univeristy of Iowa. Published, in part, in Ruch *et al: Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Co., 1926), pp. 23-53.

[2]Monroe's average reliability coefficient was higher than that found by the author and his associates, although the reliabilities were approached by somewhat different procedures, making exact comparisons impossible.

TABLE 13

Reliability Coefficients of Eight New York Regents' Examinations in the Social Studies, Year vs. Year, and Reader vs. Reader

Examination	Sub-lot	(a)*	(b)	(c)	(d)	(e)	(f)	N
Major Sequence, Course A, 1923 (Test 25) vs. 1924 (Test 26)	1	.78 ±.040	.73 ±.044	.39 ±.080	.74 ±.043	.41 ±.079	.38 ±.081	50
	2	.42 ±.078	.85 ±.026	.04 ±.095	.47 ±.074	.02 ±.095	.41 ±.080	50
Major Sequence, Course B, 1923 (Test 29) vs. 1924 (Test 30)	1	.93 ±.013	.74 ±.043	.27 ±.088	.76 ±.040	.61 ±.059	.62 ±.058	50
	2	.70 ±.048	.81 ±.034	.53 ±.060	.73 ±.045	.36 ±.083	.83 ±.031	50
	3	.68 ±.051	.67 ±.054	.48 ±.075	.46 ±.076	.60 ±.061	.45 ±.076	50
	4	.80 ±.034	.84 ±.028	.72 ±.045	.74 ±.043	.61 ±.060	.78 ±.037	50
	5	.88 ±.022	.83 ±.032	.59 ±.066	.78 ±.040	.66 ±.057	.49 ±.076	44
Major Sequence, Course C, 1923 (Test 33) vs. 1924 (Test 34)	1	.82 ±.030	.26 ±.089	.24 ±.089	.54 ±.067	.54 ±.067	.16 ±.092	50
	2	.79 ±.035	.86 ±.024	.47 ±.07	.61 ±.059	.46 ±.075	.53 ±.050	50
Civics, 1923 (Test 37) vs. 1924 (Test 38)	1	.81 ±.033	.70 ±.048	.01 ±.095	.18 ±.092	.04 ±.095	.15 ±.093	50
	2	.68 ±.051	.56 ±.065	.24 ±.090	.28 ±.087	.31 ±.086	.36 ±.094	50
	3	.68 ±.050	.56 ±.065	.27 ±.088	.04 ±.095	.11 ±.094	.10 ±.094	50
	4	.67 ±.052	.58 ±.063	.13 ±.094	.30 ±.087	.28 ±.087	.17 ±.092	50
Means of the Columns		.74	.69	.34	.51	.38	.42	
Means of the Pairs of Columns		.72		.42		.40		

*Column (a): reliability coefficient when 1923 examination was read independently by Scorer No. 1 and No. 2.
Column (b): reliability coefficient when 1924 examination was read independently by Scorer No. 1 and No. 2.
Column (c): reliability coefficient when both 1923 and 1924 examinations were read by Scorer No. 1.
Column (d): reliability coefficient when both 1923 and 1924 examinations were read by Scorer No. 2.
Column (e): reliability coefficient when Scorer No. 1 read the 1923 examination and Scorer No. 2 read the 1924 examination.
Column (f): reliability coefficient when Scorer No. 2 read the 1923 examination and Scorer No. 1 read the 1924 examination.

In an early and important study[1] Wood states: "It was to get a straight reliability coefficient on the traditional essay examination that 117 booklets were re-graded. The correlation of the first grading with the re-grading is $r = .663$." (It is to be noted that Wood obtained .905 for the reliability of an objective examination in the same subject.)

In another place Wood reports a most interesting sidelight on the reliability of the marking of papers. He says:

> The facts of a subjective scale are well illustrated in the following anecdote concerning the grading of history papers by a group of college professors of history in the summer of 1921. One of the five or six expert readers assigned to a certain group of history papers, after scoring a few, wrote out for his own convenience what he considered a model paper for the given set of ten questions. By some mischance this model fell into the hands of another reader who graded it in perfectly bona fide fashion. The mark he assigned to it was below passing, and, in accordance with the custom, this model was rated by a number of other expert readers in order to insure that it was properly marked. The marks assigned to it by these readers varied from 40 to 90.[2]

Wood, as quoted by Symonds,[3] found that in 1921 the College Entrance Board Examination in algebra (Mathematics A) had a reliability of 0.76. The geometry examination (Mathematics C) had a reliability of 0.61. Such examinations are three hours in length.

In his most recent, extensive study of old- and new-type examinations, Wood reports these coefficients of reliability for ninety-minute essay examinations in modern languages:[4]

Subject	Reliability	No. of Cases	Subject	Reliability	No. of Cases
French II	0.788	1105	Spanish II	0.722	1016
French III	0.738	867	Spanish III	0.700	629
French IV	0.415	85	Spanish IV	0.695	95

[1]From Ben D. Wood, *Measurement in Higher Education* (Yonkers-on-Hudson, New York: World Book Company, 1923), p. 193.

[2]*Educational Administration and Supervision*, Vol. VII (1921), pp. 301-304.

[3]P. M. Symonds, *Measurement in Secondary Education* (New York: The Macmillan Company, 1927), p. 297. Quoted by permission of the Macmillan Company.

[4]B. D. Wood, *New York Experiments with New-Type Modern Language Tests* (New York: The Macmillan Co., 1927), p. 115.

It should be noted that Wood's coefficients of reliability for new-type examinations in the same subjects showed a range of from 0.880 to 0.907, there being no over-lapping at all in the values of the reliability coefficients for the two types of examinations.

In a study of examinations in law[1] Wood presents further evidence on the relative reliabilities of old- and new-type tests in a college subject. After showing that traditional and objective examinations over the same course in law showed an average correlation of 0.55, Wood shows that much of this lack of correlation is explained by the unreliability of the examinations. He next attempted to determine whether the unreliability was chiefly to be attributed to either set of examinations or whether they were equally unreliable. The old-type law examinations yielded reliabilities of 0.59 to 0.73 with an average of 0.66. The reliabilities of the new-type tests ranged from 0.72 to 0.92 with an average of 0.81.

Wood's next step in the argument was to obtain measures of the validity of each type of examination by correlating each with law-school marks. To do this he compared the marks received by 215 students in pairs of law courses, with the following results (p. 7):

Pairs of Law Courses	Correlation of Essay Examination Grades With Marks in Course	Correlation of New-Type Test Grades With Marks in Course
1 and 2	0.38	0.80
1 and 3	0.41	0.74
1 and 4	0.47	0.72
4 and 2	0.47	0.78
4 and 3	0.56	0.74
Average Correlation	0.46	0.76

[1]B. D. Wood, "The Measurement of Law School Work" Part II, *Columbia Law Review*, Vol. XXV (1925), pp. 1-16.

REDUCTION OF SUBJECTIVITY THROUGH SCORING RULES

Reducing variations in teachers' marks by means of scoring rules. In Chapter I the question was raised as to the feasibility of attempting to make the traditional examination more objective without changing the fundamental nature of such examinations. It is probably possible to reduce the subjectivity of essay examinations to a considerable degree by the formulation of and strict adherence to scoring rules or schedules. In a subject like arithmetic or algebra there is always the question of how to apportion the total credit between choice of the correct solution and pure accuracy of computation. In almost any sort of examination there will arise cases where the pupil's thinking is correct, but his expression of ideas is faulty. Likewise we must face decisions as to the proper amount of penalty, if any, for carelessness, poor handwriting, errors in grammar, misspelled words, etc.

It must be true that many of these issues could be settled by rule. The rule would not necessarily be *absolutely* defensible; in fact it need not be. It may be sufficient that there be uniformity in handling the examinations of different pupils.

Two studies will be reviewed in some detail. These are probably indicative of the refinements possible through the use of scoring rules or schedules. Needless to say, the use of such guides will greatly increase the labor of marking papers, but at the same time they should increase the confidence of the teacher in her gradings.

Kelly's experiment. F. J. Kelly[1] had six fifth-grade teachers give the same examination in arithmetic to their pupils. Each teacher marked her papers but did not record

[1]F. J. Kelly, "Teachers' Marks," *Teachers College Contributions to Education*, No. 66 (New York: Columbia University, 1914), p. 83.

the marks on the papers. The superintendent had an able teacher prepare a set of rules for scoring. The teachers then regraded the papers, using the scoring rules.

Table 14 shows the two sets of gradings, the six teachers being represented by the letters A, B, C, etc.

The teacher who prepared the scoring rules also graded all the papers, using her rules. The marks of this teacher, who may be called the "judge," were used as the basis of calculating the entries in Table 14. The table is read as follows: A difference of twenty-one or more was found between the judge and two pupils marked by Teacher E when no scoring rules were used. Differences of from sixteen to twenty were found in a total of three cases, viz., by Teachers A, E, and F. The direction of the differences is also shown.

The improvement from the use of scoring rules is very marked. If we take the position that disagreements of no more than five points are not very serious, almost ninety-five per cent of the 219 pupils were marked with reasonable accuracy when rules were employed, while in the absence of rules but sixty-two per cent showed differences of five points or less. Even more striking is the fact that there were but 5.5 per cent of zero differences without scoring standards in contrast with sixty-three per cent when rules for scoring were used.

The experiment of Fauber and Ruch. O. W. Fauber,[1] under the direction of the author, repeated and extended Kelly's experiment.

Fauber's procedure was that of asking forty teachers to grade the same arithmetic paper without any specific rules. The same paper was then graded by forty-eight teachers using a carefully prepared scoring schedule. Table 15 shows the results.

[1]Unpublished M. A. Thesis (1926), University of Iowa.

TABLE 14

DISTRIBUTIONS OF DIFFERENCES BETWEEN TWO SETS OF TEACHERS' MARKS ON FIFTH-GRADE ARITHMETIC PAPERS—FIRST, WITHOUT ANY EFFORT TO UNIFY THE METHODS USED, AND SECOND, BY A COMMON STANDARD (Modified from Kelly)

	DIFFERENCES	WITHOUT SCORING RULES							WITH SCORING RULES						
		Teachers						TOTAL	Teachers						TOTAL
		A	B	C	D	E	F		A	B	C	D	E	F	
NEGATIVE	21 or more					2		2							
	16 to 20	1				1	1	3							
	15						2	2							
	14						1	1							
	13					1	2	3							
	12		1			1		2							
	11			1		1	2	4			1				1
	10						1	1		1					1
	9		1			2	1	4							
	8				1	3	1	5							
	7	1	1		1	1	1	5				1			1
	6		2			1	1	4							
	5		1	2	1	1	2	7							
	4	2	2	2	1	1	2	10	1					1	2
	3		4	2	1	2	2	11			1	1	1		3
	2	2	2	1	1	1	1	8	4	1	1	3	7	1	17
	1		5	4	3	2	4	18	2	3	4	5	1	1	16
	0	1	4	4	1	1	1	12	22	30	16	16	29	26	139
POSITIVE	1	2	5	2	2	2	1	14	5		2	2	1	3	13
	2	6	1	3	2	3	1	16	1	1	3				5
	3	9		2		2		13		2	2	1		1	6
	4	5	1	4	1	5	1	17		2	3	3			8
	5	2	3	2	2	1		10		1	1	2			4
	6	1	1		3	2		7			1	1			2
	7		1	1	6	1		9							
	8		2	1	2		1	6							
	9	1		1	2			4							
	10	1			1		1	3							
	11		1	1				2							
	12	1			1		1	3							
	13		1			1	1	3							
	14						1	1							
	15			1	1			2					1		1
	16 to 20		2					2							
	21 or more			1	3	1		5							
Totals		35	41	35	36	39	33	219	35	41	35	36	39	33	219
Medians		+3	0	+1	+6	−1	−4	+1	0	0	0	0	0	0	0

TABLE 15

FAUBER'S RESULTS ON THE SCORING OF AN EIGHTH-GRADE ARITHMETIC PAPER WITH AND WITHOUT DETAILED SCORING RULES

MARKS GIVEN	WITHOUT RULES	WITH RULES
80-84	1	
75-79	1	1
70-74	3	11
65-69	7	12
60-64	9	16
55-59	6	7
50-54	3	0
45-49	3	0
40-44	2	1
35-39	1	
30-34	1	
25-29	2	
20-24	1	
No. of Teachers	40	48
Median	60.1	64.5
Upper Quartile	65.9	69.5
Lower Quartile	49.5	60.7
Range of Middle 50%	16.4	8.8
Semi-interquartile Range	8.2	4.4
Total Range	22-84	42-77

The variability was reduced almost one half when rules were used; the semi-interquartile ranges being 8.2 and 4.4, and the total ranges being 60 and 35.

Fauber's results are in fair harmony with those of Kelly, although he did not succeed in eliminating subjectivity to a satisfactory degree when we consider that, even with detailed guides to scoring, forty-eight teachers marked the same paper all the way from 42 to 77.

In a second study Fauber took a paper in eighth-grade history which was high in content value. This paper is called Paper 1A in Table 16. He next changed this paper, re-writing it in a careless fashion, making errors of various sorts in grammar, punctuation, spelling, etc. Parts were hastily scratched out, and a few ink blots were made. The re-written paper is termed Paper 1B. No changes were made in content or thought.

TABLE 16

FAUBER'S RESULTS ON THE INFLUENCE OF FORM ON THE MARKING OF A PAPER IN U. S. HISTORY

TEACHER NO.	1	2	3	4	5	6	7	8	9	10	11	12
PAPER 1A	96	92	93	91	91	100	82	92	91	100	76	94
PAPER 1B (CONTENT)	95	92	93	91	92	100	83	92	91	100	76	94
DEDUCTIONS ON PAPER 1B:												
Neatness	10	2	5	5						4	4	1
Form		2		5						4	5	2
Spelling		2		10						4	5	2
Grammar		2		5						14	3	1
TOTAL DEDUCTIONS	10	8	5	25	10	0	10	0	0	26	16	1
FINAL GRADE ON PAPER 1B	85	84	88	66	82	100	73	92	91	74	60	88

The range of marks assigned by the twelve teachers to Paper 1A was from 76 to 100, or twenty-four points. The range for Paper 1B was from 60 to 100, or forty points (the range for content alone being from 76 to 100, or twenty-four points, as was the case for Paper 1A).

It seems clear that the range of marks is partly due to the variations in teachers' practices relative to deductions for faulty language, neatness, etc. This increase is about sixty-six per cent according to Fauber's findings if we take the variation in content of Paper 1B as twenty-four points and the variation in final marks as forty points, the additional sixteen points being the variability due to language factors.

The conclusions to be drawn from the work of Kelly and Fauber are:

1. The variability due to subjectivity of scoring in the traditional examination may be cut at least in half by the use of carefully laid-down scoring rules.

2. In spite of this fact, the traditional examination remains highly subjective.

3. Much of the variability in teachers' markings of examination papers arises from varying practices in penalizing pupils for grammatical, dictional, punctuation, spelling, and careless errors.

4. Although it can hardly be held that essay tests may be refined sufficiently by the use of scoring rules, such examinations must be employed at times, and the use of scoring rules will eliminate some of the unreliability present.

SUMMARY, DISCUSSION, AND CONCLUSIONS

A number of investigations summarized. Table 17 was assembled from a number of sources; principally the writings of Monroe, Wood, and the author and his students. This table must not be taken without reservations, since such an assemblage of correlations necessarily includes a wide variety of different situations. To be specific, widely different conditions existed with respect to such matters as: the numbers entering into the correlations, the school subject, the level of maturity of the pupils (elementary, secondary, and college), the lengths of examinations, differences in range of ability (heterogeneity), etc.

Table 17 shows a median reliability coefficient of 0.59 which is noticeably lower than the median value reported by Monroe. (See Table 9.) In view of all the facts, we must conclude that the central tendency of the old-type examination is toward a reliability not far from 0.60 to 0.65; it is certainly less than 0.70 on the average.

The investigations reviewed in this chapter are in complete harmony on all major issues. All point toward the subjectivity of teachers' marks and the older forms of examinations. If space permitted it would be interesting to comment on these studies in greater detail, but, in the

main, the tables and summaries tell their stories without need for extended comment.

TABLE 17

SUMMARY OF RELIABILITY COEFFICIENTS FOR TRADITIONAL (ESSAY-TYPE) EXAMINATIONS, AS REPORTED BY VARIOUS INVESTIGATORS

RELIABILITY COEFFICIENT	NO. OF TIMES REPORTED
.93 to .97	9
.88 to .92	12
.83 to .87	15
.78 to .82	24
.73 to .77	24
.68 to .72	19
.63 to .67	24
.58 to .62	22
.53 to .57	18
.48 to .52	13
.43 to .47	21
.38 to .42	17
.33 to .37	11
.28 to .32	11
.23 to .27	14
.18 to .22	8
.13 to .17	8
.08 to .12	3
.03 to .07	4
−.02 to .02	2
−.03 to −.07	2
−.08 to −.12	0
−.13 to −.17	1
−.18 to −.22	2
−.23 to −.27	1
Median=.59	Total 285

We must return to the list of objections with which this chapter opened. The first two have previously been discussed at length. It is only necessary to state once more that the traditional examination is open to two serious limitations:

1. Subjectivity of marking, and
2. Limited sampling.

Limited sampling is unavoidable in any examination, new-type or old-type. A comparison of the older examination with the newer on the score of sampling would seem at first sight to present little ground for choice. This is not quite true for two reasons:

1. As was shown in Chapter II, the two sorts of examinations differ in their theories of sampling. The traditional examination employs an *intensive* type of sampling; the new-type, an *extensive* sampling. There is considerable logical (and some experimental) evidence of the superiority of the extensive sample. (See Chapter II.)
2. The new-type examination can cover far more ground in the same amount of working time because there is no need to spend time in writing a mass of words. The response by underlining, encircling, checking, etc., is so rapid that at least ninety per cent of the examination period is spent in thinking about the responses. With the traditional examination a larger fraction of the time is spent in writing.

Excessive writing of answers is uneconomical. Prominent among the objections to the traditional examination is the charge that it is wasteful of the pupil's time. In a sixty-minute examination a pupil spends from fifteen to thirty (sometimes more) minutes in writing his answers. If no writing were necessary, the examination might include at least twice as many questions. Worse still, most of the words which he writes convey little information about his real knowledge of the subject. Language requires a large number of "filler words," useful to be sure for grammatical reasons, but useless for examination purposes if we view the examination as a measuring instrument.

Traditional examinations encourage bluffing. This charge must be admitted, although something analogous, and perhaps fully as objectionable, is inherent in many of the new-

type tests. The reference is to the opportunity for guessing correct answers in such tests as the true-false and multiple-choice.

The phenomenon of bluffing needs no comment. It exists in all examination situations. The nature of the essay examination makes it open to such abuses. Pupils know with a somewhat uncanny intuition that teachers are loath to mark any question zero unless the question is left absolutely blank. There is an ever-present, and perfectly human and philanthropic, tendency to reward effort, no matter how misguided and futile. Laudable as are such "weaknesses of the flesh," they do not serve the ends of measurement. It will be recalled that four teachers out of eighty-nine rated as better than zero the statement, "The five largest cities of the United States are (1) navada (2) Arkansas . . . " in reply to the question, "Name and locate five of the largest cities of the United States, and name their leading industries, exports, and imports."

There is a genuine difference between old- and new-type examinations in one respect. When a pupil is confronted with a broad discussion question, he in one sense chooses the line of attack. He may be entirely ignorant of the import of the question, but for the time being he is the general in charge. He can naïvely "misunderstand" the question and write on some alien topic where his meager store of knowledge can be turned to better advantage. He can at times go around, under, or over the topic in a very skillful manner. He has nothing to lose, and he might win in the hands of a philanthropic teacher. An objective test, on the contrary, forces him to "face the music." In this case the teacher chooses the battleground. The examination forces the pupil to react to those things which the teacher deems important. It gives her, as it should by right of more mature wisdom and judgment, the leadership in the examination period.

On the whole, exchanging the disturbing factor of bluffing for the admitted danger of guessing (in many new-type tests) is a gain, since there is no mathematical formula for minimizing bluffing but there is a more or less adequate statistical means of allowing for guessing.

The answer to a question of the discussion type is a complex thing. It is made up of many sorts of elements. To mention a few:

1. There are some statements which are true, and to the point.
2. There are other statements which are true, and beside the point.
3. There are statements which are absolutely false.
4. There are many things which the teacher hoped would appear in the answer, but which are missing.
5. There are half-truths, garbled statements, and ambiguous statements which reduce in extreme cases to meaningless sequences of words.

How can such diverse elements be fused into a single judgment? The answer seems to be that they cannot; at least the weight of the evidence of this chapter is that they cannot be evaluated with any high degree of accuracy.

Final conclusions. The element of subjectivity in the traditional examination is a source of marked unreliability. Such examinations have been found to be wasteful of time in the sense that excessive writing of words which convey no knowledge of accomplishment to the marker of the paper is required. Such wasted time would allow the answering of a much larger number of short-answer or objective questions. An objective test over the same ground covered by a traditional examination would yield a far more extensive sampling of the pupil's knowledge. Subjectivity of marking may be reduced about one-half by the adoption of and adherence to a set of scoring rules when essay examinations are to be

graded. Such scoring rules increase the labor of scoring papers, but are nevertheless highly desirable. The traditional examination should be employed principally when the subject-matter does not lend itself to completely objective measurement; even in such cases the results must be taken with a great deal of conservatism. A combination of traditional and new-type examinations should probably be used in many school subjects, especially where present knowledge is unable to suggest purely objective types of measurement.

CHAPTER IV

ADVANTAGES AND LIMITATIONS OF OBJECTIVE EXAMINATIONS

General statement. The general course of argument presented in this chapter may be made clear by this outline:

I. Advantages of objective examinations
   1. Objectivity (freedom from personal opinion) in scoring
   2. Extensive sampling
   3. High reliability per unit of working time
   4. Economy of scoring
   5. Freedom from bluffing
   6. Greater control of the examination system by the teacher

II. Limitations of objective examinations
   1. No provision for language training
   2. Open to guessing and chance
   3. Reputed to measure only factual memory
   4. Said to be an unnatural method of using school-acquired information
   5. Test recognition rather than spontaneous recall

Many of the above points have been commented on at some length in preceding chapters. The present chapter is therefore somewhat of a summary and co-ordination of points of view.

ADVANTAGES OF THE OBJECTIVE EXAMINATION

Objectivity of scoring. Objectivity of scoring needs little further comment. Chapter III has shown the dangers of

subjective scoring. When 100 (or more) teachers grade the same paper all the way from 28 to 92 (as Starch found for a geometry paper), or when they evaluate the same question so differently as from 3 to 20 (as Ruch found for a geography paper), the only conclusion which can be drawn is that such examinations do not *measure* the pupils.

Objectivity is a prime essential for reliability of measurement. Much of the educational process must be highly subjective, but measurement implies accuracy. The examination should exclude so far as is possible the personal opinions, biases, whims, and temperaments of teachers. The examination should be a measure of the pupil, unadulterated by factors which represent the psychological reactions of the teacher.

The objective or new-type examination can be made wholly or almost wholly objective. The traditional examination cannot. It follows that a high degree of objectivity can only be had through the new-type test. To the extent to which this is true, the objective test has a clean-cut advantage over the older forms of examinations.

Of the two principal sources of unreliability in examinations (subjectivity and limited sampling), the former only is completely eliminable. It follows, therefore, that this source of inaccuracies in examination marks should be minimized or eliminated completely.

Thomson mentions the question, "Describe the universe and give two examples" as the extreme form of such subjective questions.[1] This question may not be real, but it has as its "twin" such a question as "What were the contributions of Babylonia to civilization?" or, "Describe the geography of Greece."[2] It is only fair to contrast such questions with the misguided attempt at objectivity reported by

[1] G. H. Thomson, *Instinct, Intelligence, and Character* (New York: Longmans, Green and Co., 1925), p. 202.

[2] *See* W. J. Osburn, *Are We Making Good at Teaching History?* (Bloomington, Illinois: Public School Publishing Co., 1926), for tabulations of the kinds of questions which history teachers actually ask.

Brinkley,[1] who found the following: "The ..... who ..... was ..... ." The expected answers (it may be interesting to learn) were "man," "discovered America," and "Columbus."

Extensive sampling. There can be no real dispute on this point. The ordinary five- or ten-question discussion type of examination requires a great deal of writing. It follows that a large fraction of the words actually employed in phrasing an answer to such a question are "filler" words, i. e., words needed to complete the sentence structure but valueless in adjudging the merit of the pupil's answer. The word counts of Thorndike and Horn have shown that the most frequently recurring words in the written language are: *a, about, all, also, am, an, and, any, are, as, at, be, been, by, can, do, for, get, has, have, he, I, if, in, is,* etc. None of these words, ordinarily, may be expected to convey any knowledge about the child's achievement.

Rough studies by the author of the answers to questions given by pupils on hundreds of uniform state eighth-grade examinations show that from four to ten lines of written answers are given to each question. When these answers were analyzed, the central tendency seemed to be that from four to seven different ideas or facts were reported. This means that the effective length of a ten-question examination was from forty to seventy items. Such examinations required (or were allowed), usually, sixty minutes. The sampling *per minute of working time could hardly have been more than one idea (or fact) per minute.* This is very small indeed. Experiments reported in Chapter XI show that sixty minutes of objective testing would have permitted at least three times as extensive an examination.

The fact is that the usual discussion or essay examination is very wasteful of the pupil's time. Half to three-fourths

[1] S. G. Brinkley, *Values of the New-Type Examinations in the High School* (New York: Columbia University, *Teachers College Contributions to Education,* No. 161, 1924), p. 36.

of the examination period is spent in writing words, useful enough for purposes of sentence structure but quite valueless in conveying to the teacher any facts about the pupil's knowledge of the subject. By checking, underlining, or inserting occasional words, from three-fourths to nine-tenths of the writing may be eliminated in an examination. This results in an obvious economy of effort, an increased allowance of time for thought, and an increased extensity of sampling in a given amount of examination time.

As has been said repeatedly in this volume, reliability depends principally on two factors, objectivity and extent of sampling. The unreliability arising from subjectivity may be eliminated completely. Unreliability due to sampling cannot. It will always be present since measurement is invariably a sampling; never complete. The key to reliable educational measurement is therefore extensive sampling in order to secure an accurate and stable measure of each pupil.

In Chapter II it was shown that the underlying theory of sampling was different in the old and the new types of examinations. The former represented an *intensive* coverage of a very small number of topics; the latter represents a more *extensive* sampling of topics but a less thorough coverage of any one topic. When time is limited, as is usually the case with examinations, somewhat more reliable results are to be expected from what the author has termed the *extensive* sampling.

High reliability per unit of working time. There is nothing of importance to add to the discussion of the two preceding sections. Economy demands that no unnecessary writing be done during the examination. If discussional types of tests are to be given, we must face frankly the situation of giving two to three times as much time to examinations as is required by the newer and more economical objective tests. The reliability of the ordinary ten-question essay examination of sixty minutes can hardly be held to

average more than 0.60 to 0.70 in view of the available evidence. In later chapters of this book (especially Chapter XI) it will be shown that sixty-minute objective tests may show reliabilities as high as from 0.70 to 0.90, provided care is used in formulating such tests. The sixty-minute objective test may contain, not five or ten questions, but as many as from 100 to 200 items.

The author once studied a series of written examinations as administered by a state department of public instruction to all eighth-grade pupils. He selected those subjects which are covered by a well-known battery of standard educational tests, and determined the reliability of this composite. By the usual prediction formula devised by Spearman and Brown it was shown that in order to equal the reliability of the standard test (of about two and one-quarter hours working time), it would be necessary to do three and one-half weeks of testing by the type of examination actually employed. This may be an extreme case, but it is a good illustration. The name of the state is withheld, although no especial reflection upon that state is implied, other than upon the continued adherence to such an examination.

We dare not close this section without reference to one experimental study which is not entirely in harmony with our point of view. Brinkley, (comparing old- and new-type examinations) found that "given tests of equal length, as measured by time spent in testing, and prepared by teachers with some training in the matter of test construction, one type of test yielded practically as good results as another for measuring achievement in history in the senior high school."[1]

This study of Brinkley's is somewhat disconcerting and difficult of explanation. It certainly stands as an exception to the findings of such investigators as Wood, Toops, De-

[1]S. G. Brinkley, *Values of New-Type Examinations in the High School* (New York: Columbia University, *Teachers College Contributions to Education*, No. 161, 1924), p. 58.

Graff, Stoddard, and Ruch. Symonds, after summarizing and reviewing the earlier literature on objective tests (including the work of Brinkley), says:

> The evidence is decisive that new-type examinations are more reliable than traditional examinations. In fact, the evidence is more decisive than an off-hand inspection of the above tables [not reproduced here] would indicate. One usually wishes to compare examinations for reliability under equal working times. Monroe does not state the average length in time of the sixty-six sets of examinations used by him in getting his reliability coefficients. The median number of questions in an examination was seven, and one may assume that the examination time was at least an hour . . . Ruch's correlation of .896 for 100 recall questions was based upon 18.7 minutes of testing time. Thus it is a matter of comparing an average reliability of .65 for the traditional examination obtained in sixty minutes, more or less, of testing time with a reliability of .90 for a new-type test obtained in nineteen minutes of testing time. The superiority of the new-type test, so far as reliability goes, is plainly evident.[1]

Brinkley himself states that "the differences were not in any of the cases large." The new-type tests made by classroom teachers after some preliminary training were not quite equal in validity to the old-type tests, but were better than those made before this training was given. On the other hand, the tests made by Brinkley himself were not inferior to the traditional examinations, but, on the whole, were somewhat superior.

The experimental group used by Brinkley was not large. Starting with 163 pupils, only ninety-five sets of records were finally available. This tended to obscure the significance of any differences noted. It is also to be remembered that each of the ten tests used contained but thirty-three items, divided as follows:

True-false	100	Word or Phrase Answer	52
Multiple Choice	89	Arrangement	20
Completion	66	Essay	9

[1] P. M. Symonds, *Measurement in Secondary Education* (New York: The Macmillan Co., 1926), p. 298. Quoted by permission of the Macmillan Co.

Moreover, the plan of scoring the essay examinations was one which is not typical of prevailing practice, and one certain to improve the reliability and validity of its results. Brinkley says:

> The scoring of the essay examinations was rendered as consistent as possible by listing beforehand the items that should be included in a correct answer or the comparison that should be made. Often a set of papers was read and additional items entered on the key as a result before the real scoring was begun. Values were then assigned to the different items so as to give a total score of 10 for each question, and the papers scored by comparison with this key.[1]

As has been said, this plan of scoring is not typical of the usual practices, and was one certain to improve the reliabilities of the essay examinations. This study by Brinkley, although carefully executed, should be repeated with larger numbers, and perhaps under conditions somewhat more typical of actual practices, before it should be held to refute a number of more extensive and equally careful experiments which have yielded quite different results.

Economy of scoring. There is no dispute on this point. Objective tests if planned carefully as to mechanical features may be scored from two to five or more times as rapidly as can essay tests of comparable length. The term "comparable" has been used advisedly since there are a number of bases for such comparisons. We might compare old- and new-type tests of equal lengths in terms of (1) numbers of items, or questions, (2) equal working times, or (3) equal reliabilities.

The first basis is, of course, quite indefensible since an objective test item is quite a different unit from an essay question. To compare tests of the two types having equal

[1]See Chapter III for evidence that subjectivity might well have been reduced by one-half by Brinkley's use of carefully drafted rules for scoring. Brinkley does not seem to give adequate weight to this fact when he concludes that old- and new-type examinations are roughly equal in merit. A fairer comparison would have been to measure objective tests against essay tests read by the prevailing methods actually in use by teachers, i. e., without detailed scoring rules.

working times is a better basis, but, as has been shown, the objective test is *in effect* a longer test under such a condition since it is more economical of time. On the whole, it is more accurate to think of the relative economies in terms of tests which are equally reliable. This will mean, in general, that a ten-question essay examination is to be compared with objective tests of perhaps forty to seventy-five items. The latter can be scored at rates of twenty to thirty or more per hour, depending upon the type of test and the mechanical arrangement of the responses, etc. Essay examinations of ten questions cannot be scored, as a rule, faster than from ten to fifteen an hour.

Freedom from bluffing. This advantage has already been pointed out (Chapter III). It has its analogue in the so-called guessing element in many objective tests. A detailed consideration of guessing is reserved for Chapter XII. Bluffing is even less desirable than guessing, since partial control of the latter is possible through mathematical formulas.

Greater control of the examination by the teacher. So far as the author is aware, this superiority of the objective examination has not been commented upon by other writers.

It is well known that a pupil who is shaky on the answer to an essay question can often "get by" by pretending to misunderstand the intent of the question, and then writing on something more to his liking and information. To some extent, therefore, the pupil chooses his reactions to such a question. The question may encourage this in two ways: (*a*) by ambiguity of statement, or (*b*) by intention, i. e., some teachers prefer to give the pupil great latitude in choosing the direction of his replies. Desirable as the latter may be in theory, it makes for greater subjectivity in evaluating responses.

With the new-type examination *the teacher forces the pupil to react, one way or another, to what she thinks is important.* (The items of the test are supposedly her concept of what is of prime importance.) There can be no doubt that this practice extends the teacher's control over the examination situation. Moreover, the teacher has exactly the same right to control the nature of the pupil's reactions during an examination as she has to select subject-matter, methods, etc., in the larger phases of instruction. Objective examinations cannot "misfire" to the same extent as essay tests which so often prove unsatisfactory because the pupils did not succeed in writing on the issues intended by the teacher.

LIMITATIONS OF THE OBJECTIVE EXAMINATION

Neglect of language training. This admitted limitation of the new-type test was discussed in Chapter III, where the claims of the traditional examination were also considered and criticized. Suggestions were made as to methods of attaining increased language facility by written examinations.

The "guessing" element in objective tests. This matter has also received passing attention in preceding chapters. Chapter XII of Part III will take up in detail the history of the controversy on the degree of invalidation of true-false and multiple-response tests resulting from the large amounts of guessing possible.[1]

Objective examinations measure memory only. This is a charge which is easily made by opponents of the new-type examination. Moreover, this charge too often makes the tacit assumption that the traditional examination is free (or at least, freer) from this criticism.

[1]The author cannot forebear quoting a somewhat incoherent communication recently received from a teacher. Speaking of true-false tests, she said, in part (italics mine): "The element of guesswork is strong and the *gamboling instinct* is brought to bear. If they (the pupils?) are penalized, i.e., 'miss one, charge two' is worse still. It is dishonest."

If there is any one place in examination practices where loose thinking prevails, it is relative to the so-called "thought" question. Teachers and educators pay lip service to the thought question and then proceed merrily to ask pupils to "Name the principal products of New England" or to "List the main causes of the Revolutionary War." The head master of a famous military school once wrote the author to the effect that he had taken a six-weeks' course in one of the largest universities for the training of teachers. For six weeks the professor waxed warm and loud in his praises of the thought question—and never once gave a concrete illustration!

It is not the intention either to deny the existence of the thought question or to decry its merits. It is merely suggested that we clarify our ideas on the point before attempting criticism of any type of examination. Perhaps it behooves the author to come forth with good examples of thought questions at this point. This challenge will not be met fully for many reasons, not to mention possible inability.

Suppose we take as a fair approximation to a thought question, "Why do many taxation authorities believe that the income tax is the fairest form of taxation yet developed?" The average man, whose experience with taxation consists in little more than semi-annual grumbling when taxes fall due, might, if intelligent, reason out certain valid arguments in favor of this proposition. To the extent that he did this, the question is a thought question. On the other hand, a high-school senior or junior, having just completed a course on economics, might answer in one-two-three order the points concisely summarized and presented on page 289 of his textbook. The *same* question is a matter of *pure memory* in this case.

The point is: *The difference between a thought and a memory question does not reside principally in the question itself but in the mental background of the pupil.* Paul Smith, an intelli-

gent but indolent boy, may sit next to George Brown, of moderate ability but good memory, and receive the same instruction and write the same examinations. Paul may do a "lot" of thinking, much of it original, when examination day comes. George may merely place on paper his carefully hoarded information. George's paper may be superior by far, but Paul did whatever thinking was called forth by the thought questions asked.

In a sense the difference between thought and memory questions is chronological. The question in mathematics which is an "original" for the eighth-grade pupil reduces to sheer memory by the end of his high-school course.

As an approximation to a good thought question, the following is submitted:

State what you might have prophesied as to the future of the Roman republic, if you had lived in the first century before Christ and had known the following facts: Marius becomes consul for the seventh time; Sulla is given the title of "Perpetual Dictator"; Caesar becomes dictator for life. State additional facts that support your conclusion.[1]

The foregoing is not an easy task for the average high-school graduate. Compare it with the following attempt to objectify the same question.

Directions: Check (X) the statement which expresses what you might have prophesied as to the future of the Roman republic if you had lived during the first century before Christ and had known the following facts:

Marius becomes consul for the seventh time.
Sulla is given the title of "Perpetual Dictator."
Caesar becomes dictator for life.

1. ..... The republic was on the verge of developing a greater democracy.
2. ..... The army becomes less aristocratic, and Marius enlists all men who desire to fight.
3. ..... The senate desired to grant greater economic rights to the working class of Rome.

[1]Taken from an examination of the New York State Regents in History, Major Sequence, Course A, 229th High School Examination, June 20, 1923.

4. ..... Civil wars and the military rule of one man power would in time overthrow the republic.
5. ..... The rule of the assembly and its leaders was about to triumph over the rule of the senate.

Directions: Check two (2) additional facts which support your conclusion as indicated above.

1. ..... Rome became a great manufacturing city, thus changing the political organization.
2. ..... The Italians wanted citizenship because the municipal government could not rule successfully the entire peninsula.
3. ..... The wealth brought in as booty from foreign wars greatly lessened the taxes which the poor were forced to pay.
4. ..... Rome would not pass laws for the relief of the poor, causing many to die of starvation.
5. ..... The women of Rome enjoyed greater freedom than the women of Greece.
6. ..... Rome conquered practically all of the known civilized world.
7. ..... Greek culture was acknowledged to be superior to Roman.
8. ..... Farmers of Italy, unable to earn a living, came to Rome in great numbers in search of work; the resulting unemployment created an economic and political problem leading to the passage of the Free Corn Laws.
9. ..... Rome lost interest in education and culture.
10. ..... The Roman Empire required all people to worship the emperor.

The objective version of this thought question is not submitted as a "model" question. The framer of this question found his task a difficult one because he was not at all certain what the original question called for. The objective form of the question does have suggestive value as to the technique of objective thought questions. It is an open question which form of the question is more valid and reliable; our experimental evidence favored the objective question in spite of whatever faults it may have. There is small chance of earning a very high score by guessing, and the scoring is about as simple as can be imagined.

It should be noted that any desired degree of discrimination of thought can be had by careful framing of the state-

ments from which the pupil must choose. In such a question some statements should have high merit, some be true but hardly pertinent, others should be quite beside the point, and some may well be entirely false.

While we are considering this particular question, we may as well scrutinize the other questions from the same examination. The reader must decide which are the thought questions, if any.

Write on architecture and building among the early Egyptians, touching on (*a*) kinds of structures built, (*b*) general plan and appearance, and (*c*) materials used.

What geographic features of Greece favored (*a*) the growth of small states, and (*b*) commerce. Explain fully in each case.

Mention *two* valuable contributions to civilization resulting from the barbarian invasion of the Roman Empire.

Describe an attack on a medieval castle, pointing out the difficulties to be overcome and the means used to accomplish the downfall of the defenders.

Write on the Magna Charta, touching on (*a*) the circumstances under which it was granted, (*b*) three provisions, and (*c*) its importance in history.

In what *three* parts of the world were England and France rivals during the 18th century? Which country gained territory as a result of this struggle? Where was this territory located?

Write in detail of the services that Peter the Great rendered to Russia.

Wood, apropos of modern foreign language tests, has made some very pointed remarks on this question of memorizing vs. thinking:

There are few teachers who do not admit that the objective tests are better as *measuring* devices, but some teachers fear that the objective tests are pedagogically unsound, and that they will tend to mechanize teaching and produce what is called "dead uniformity." Specifically, it is feared by some that the objective passive vocabulary tests will cause students, aided and abetted by their teachers, to "memorize *mere* lists of words." Thus it is feared that students might make a serious breach in future objective tests by the simple and effective device of *memorizing* the small matter of the two or three thousand most frequently used words in each language! The reader can judge for himself whether such a result would be unpedagogical

or calamitous, or an unhoped-for blessing. It may well be that the objective vocabulary test may produce such a miracle; but the proponents of new-type tests have never been optimistic enough to hope that this charge against their tests would turn out to be true.[1]

There are a number of similarly penetrating refutations of current prejudices against the new-type examination in this recent monograph by Wood. The teacher of languages, especially, will wish to read the whole discussion. The notion that pupils will study isolated facts as a preparation for objective tests is a matter closely related to the thought question since it is the type of question (not its subjective or objective form) which determines the attitude of the pupil.

Another angle of the thought question has ordinarily escaped attention, viz., that once a situation has been used for the purposes of evoking thought, it is almost valueless (as a thought question) for the same pupils at any later time. At a future date the original thought question will be answered mostly from memory. A thought question must thus be an "original" in the sense that facts common to all pupils must be organized into a reaction to a novel situation. Certain it is that in most cases we cannot be sure of what is thought and what is memory, either from inspection of the question or from scrutiny of the pupil's reply.

Returning to the two versions of the question in Roman history, what right has any one to say in advance of actual experimentation that one version is any more conducive to the evil of cramming than the other?

The framing of thought questions is an art—and a rare art so far as the author's experience goes.[2] Thought questions do exist; certainly they are valuable; but how

[1]B. D. Wood, *New York Experiments with New-Type Modern Language Tests* (New York: The Macmillan Co., 1927), pp. 96-97. Quoted by permission of the Macmillan Co.

[2]The author has for years been using as an assignment in certain classes the preparation of ten thought questions. Fully two-thirds (usually more) of the questions submitted by classroom teachers bear little or no resemblance to genuine thought questions. Typical examples are: "What are the products of New England?" "What causes fainting?" "How does a siphon work?" "Why did Columbus think the world was round?" "Why is tobacco injurious to children?" etc.

can we guarantee that they evoke thought simply because they are expressed in a form which makes a thoughtful response possible?

Psychologically there is a considerable degree of interrelationship between memories and thoughts, between ideas and facts. We reason with facts. We cannot reason correctly without them. Let us admit that there exists a problem relative to the framing and use of thought questions and refrain from assuming, a priori, that the traditional examination occupies a favored position in this respect. It may be true, but there is no proof for such a belief. It is both more scientific and more stimulating to assume that the new-type examination has great possibilities in this direction, if for no other reason than its greater objectivity. It is futile to hold to a form of question, whether it be thought- or fact-provoking, if it cannot be evaluated fairly.

Objective examinations require mimeographing or printing. This is a practical objection which must be admitted. Some kinds of objective tests (the true-false, particularly) may be read aloud to the pupils. This may work fairly well with older children, but it is always a second-rate procedure. Mimeographing takes time, and stencils cost considerable money.[1] Paper costs are roughly the same, as the answers must be written in any case. If there is any decided superiority of the new-type examination, school budgets must be made to carry the added expense. The total costs appear infinitesimal in comparison with modern school expenditures for (often) less essential (but perhaps more ornamental) equipment than mimeographs and mimeograph paper.

Objective tests are unnatural and unpedagogical. This criticism is another which falls glibly from the tongue of the critic. Wood has already disposed of one angle of the matter. The author would like to raise the question as to

[1]Some teachers find a hectograph effective and relatively inexpensive in making copies of objective tests. Forty or fifty legible copies can be made from a single stencil.

what is the "natural" or "pedagogical" form of questioning.

It would be most instructive and valuable to study the exact forms in which school-acquired knowledge is used in actual life. These modes would then be our answer as to what are the natural or pedagogical forms of questions. As a rough substitute for such experimental investigations, the author has made a conscious effort to observe the ways in which adults use their concepts and information. One fruitful source of hints has been the conversations in the smoking compartments of Pullman cars. Thus far nothing has been found which resembles closely any variety of school question-and-answer with the possible exception of a sort of true-false statement. Certainly no person after leaving the public schools has ever been called upon to write or to recite answers to such a question as "Give the causes of the Revolutionary war" or "Explain in full the digestion of carbohydrates." We do hear adults make statements to the effect that "the League of Nations is a flat failure" or that "the Monroe Doctrine is nothing more nor less than a 'chip on America's shoulder.'" To such an assertion (in the smoking compartment) there is speedy and dogmatic challenging, attempted support and refutation, and a final settling down to argument. The mental set in the ordinary political, economic, or social argument is as near to the true-false attitude as to anything which goes on in school. Whether it be the riding qualities of the new Ford, the payment of European debts, or the batting of Babe Ruth, nothing closely akin to the question of the traditional examination emerges.

We are densely ignorant of what is "natural" as a type of examination question. If naturalness is a prime desideratum of examining, the solution must come from an analysis of adult needs and usages; *not from assumptions that the old is natural and pedagogical, but that the new is lacking in such values.*

As has been stressed before, *the examination should be nothing more nor less than a sampling of the kinds of activities which go on daily in the classroom.* It would be most disappointing to believe that the traditional written question-and-answer represented the heights to which pedagogy has ascended to date. No type of examination is very natural in the sense of social utility. The "best" examination must be, for the present at least, that one which by experiment can be shown to be the most valid and reliable. Such criteria can be made to rest upon a far more sound basis than can the cries of "mechanization," "unpedagogical," "dead uniformity," "pure memorization," etc. It is a queer sort of teacher who can instruct pupils nineteen days out of a month in such a way as to avoid Scylla, but on the twentieth, or examination day, steer directly at Charybdis.

Objective tests measure recognition rather than spontaneous recall. Many types of objective tests call for selection among stated alternatives. The traditional examination called for spontaneous recall. This difference has been deemed a superiority of the older examination. The question is a difficult one to decide. It is something like the issue discussed in the preceding section. We do not know what knowledge should always be at our finger tips and what knowledge will suffice if it enables us to select truth from error when both are presented. We probably need both kinds of knowledge. The author prefers to leave this question an open one in the mind of the reader. Certain of the objective-test devices (the simple-recall and the completion) resemble greatly the traditional examination. Others (true-false, multiple-choice, matching, etc.) are recognitive or selective operations. Very little of the acquisitions of the schoolroom need be retained in a form capable of immediate recall. For many purposes it is sufficient that the right facts, ideas, and concepts "come to

mind" when choice between alternatives is presented. On the whole, there seems to be no certain knowledge upon which a just criticism of objective examinations can be made on the score of recognition vs. recall. Here again is a promising, although very difficult, field for investigation.

CHAPTER V

STUDENTS' ATTITUDES TOWARD EXAMINATIONS

Introductory statement. It is probably agreed that some thought should be given to the attitudes of both pupils and teachers toward examinations. Criticisms from either source may at times suggest changes in methods of educational measurement, provided, of course, that such criticisms are constructive in character. Both teachers and students may be expected to be biased on the issue of examinations vs. no examinations, and these biases may run in opposite directions. Pupils often dislike examinations merely because they represent irksome labor. If the examination or test is viewed as part of the educational process, we need not abandon it because of lack of popularity with pupils. Changing a punctured tire is a nuisance and extremely unpleasant, yet few would conclude that the solution is the abandonment of automobile transportation, or even the running of cars on a flat tire. Most persons agree that the tire should be fixed regardless of transitory feelings. Examinations can certainly be made more pleasurable except for the minority who resent any "showdown" on their knowledge or lack of knowledge.

This chapter will present brief abstracts of several studies of students' attitudes toward examinations, together with a short summary of these investigations. No space will or need be devoted to the non-experimental literature (and it is somewhat voluminous) on the continuance or rejection of the examination system. The *ipse dixit* arguments are principally based upon the type of examination which appears to be passing in large measure.

Somers's study of the attitudes of students toward examinations. Somers[1] has verified the common observation that pupils generally have marked aversion to the taking of both oral and written examinations. He asked forty-five teachers and 163 college students to rank their attitudes on eighteen activities, examinations included, on three scales, as follows: pleasant-unpleasant, valuable-worthless, and moral-immoral. The attitudes of students and teachers are surprisingly similar.

TABLE 18

RESULTS OF SOMERS AND GALLAGHER AND RUCH FOR EIGHTEEN ACTIVITIES ON A SCALE OF PLEASANT-UNPLEASANT

ACTIVITY	(1) RANK GIVEN BY TEACHERS (SOMERS)	(2) RANK GIVEN BY STUDENTS (SOMERS)	(3) RANK GIVEN BY STUDENTS (GALLAGHER AND RUCH)
Attending a concert	1	2	5
Reading a story	2	1	2
Attending a movie	3	4	4
Witnessing a basketball game	4	3	3
Going on a field trip or excursion	5	7	9
Attending classes	6	6	7
Going to a circus	7	9	6
Attending a convocation or assembly	8	8	8
Laboratory experiments	9	11	12
Writing a story	10	10	11
Attending a dance	11	5	1
Cleaning your room	12	12	10
Washing dishes	13	13	14
Writing a theme or composition	14	14	13
Taking an examination (state or discuss type)	15	16	15
Taking an oral quiz	16	15	16
Pulling weeds	17	17	17
Digging a ditch	18	18	18
Numbers	45	163	166

The author repeated the major part of Somers's study upon a group of 166 students and teachers, about one-fifth

[1]Grover T. Somers, "Students' Attitude Toward Examinations," *Bulletin of the School of Education*, Indiana University, Vol. III, No. 1 (September, 1926), pp. 1-48.

of the group being experienced teachers. The students comprised juniors, seniors, and graduates in roughly equal numbers. Mr. Edward D. Gallagher summarized these data, using the procedure which Somers had employed in his study.

Table 18 shows the ratings upon the Pleasant-Unpleasant scale. Columns (1) and (2) give Somers's findings and column (3) shows the results obtained by Gallagher and Ruch.

Teachers and students seem agreed that the taking of examinations is little more satisfying than weed-pulling and ditch-digging!

The correlations between the rankings given in the three columns are all reasonably high, as follows: (1) $r_{12}=0.94$ (2) $r_{13}=0.84$ (3) $r_{23}=0.96$

A second part of Somers's investigation showed, however, that a better attitude toward examinations (at least those of the objective type) resulted from a semester's use of examinations as an integral and constructive part of instruction. The increase in favorableness of attitude, although not so great as we might wish, was nevertheless significant. The final rankings of the eighteen activities in Somers's experimental groups gave the various types of examinations positions as follows: true-false test, 10; matching test, 11; multiple-choice test, 12; oral quiz, 14; and the written examination and completion test tied for 15th and 16th ranks.

One final point should be noted about Somers's study. The correlations between students' and teachers' attitudes were surprisingly high except in one case—the purposes and functions of written examinations. This discrepancy will be made clear from the data presented at the top of the next page.

Scale	Correlation
Pleasant-unpleasant	0.94
Moral-immoral	0.99
Valuable-worthless	0.91
Purpose and function of:	
Recitations	0.95
Oral quizzes	0.85
Written examinations	0.39[1] (0.61)

It seems to be true that students and teachers differ greatly as to their notions of the purposes and functions of written examinations. The rank-orders are given in Table 19 for Somers's study and its repetition by Gallagher and Ruch.

TABLE 19

Purposes and Functions of Written Examinations

Purpose or Function	(1) Rank Given by Teachers (Somers)	(2) Rank Given by Students (Somers)	(3) Rank Given by Students (Gallagher and Ruch)
Means of measuring ability to think	1	3	3
Basis for comparing achievement	2	4	5
Basis for measuring students' knowledge	3	1	1
Means of stimulating reviewing	4	8	2
Means of stimulating studying	5	7	4
Opportunity for self-expression, etc.	6	9	9
Means of measuring results of teaching	7	6	7
Means of stressing important facts	8	2	6
Basis for term marks	9	5	8
Means of measuring memory ability	10	10	10
Numbers	45	163	177

The intercorrelations of the three columns of Table 19 are: $r_{12} = 0.61$ (0.39 as given by Somers) $r_{13} = 0.78$ $r_{23} = 0.56$

All in all, Somers's conclusions that teachers and pupils hold substantially the same views on the pleasurableness of examinations but differ widely on the functions of examinations seem to be valid.

[1]This value is in error as given by Somers. According to the author's calculations it should be 0.61.

Hughes's study of students' attitudes toward examinations. Hughes,[1] in an investigation somewhat along the lines of that of Somers, reaches rather different conclusions. He gave to classes in "Problems of Democracy" examinations made up of the following parts:

PART I. Limited recall (controlled answers but not completely objective)
PART II. Specific definition or explanation (i. e., state meaning, define, or explain laws, terms, etc.)
PART III. Completion exercises
PART IV. Multiple-response items
PART V. True-false (called by Hughes, "Alternate-Response")
PART VI. Essay examination

He asked 157 pupils to record their attitudes toward these different types of questions upon a scale of seven points. The results are shown in Table 20.

Point *A* shows that the essay test suffers badly in comparison with the newer examination so far as pupils' enjoyment of the examination period is concerned. Points *B*, *C*, and *G*, which deal with the measurement aspects of examinations, indicate that the objective tests inspire greater confidence as to their justice and accuracy; it is further to be noted that the combination of types is most to be desired. The essay examination is almost "out of the running" on these points. The results on Point *G* (which is always dear to the hearts of pupils) are particularly significant. Perhaps the worst showing of the essay examination is on Point *D*, where strong feeling is evidenced that the essay examination encourages pure rote memorization of materials. Points *E* and *F* are not commented on, as Hughes felt that the pupils did not fully understand the significance of these two issues, thus giving rise to apparent contradictions with Point *G*.

[1]Unpublished Master's thesis, University of Pittsburgh. For a résumé, *see* C. A. Buckner and R. O. Hughes, "Testing Results in the Social Studies," *Journal of the School of Education*, University of Pittsburgh, Vol. I, No. 1 (Sept.-Oct., 1925), pp. 5-11.

TABLE 20

EVALUATION OF ATTITUDES OF 157 PUPILS TOWARD DIFFERENT TYPES OF EXAMINATION

POINT	PARTS I, II, AND III	PARTS I TO V	PARTS I AND II	ESSAY TYPE	NO CHOICE
(*A*) Type most enjoyed	58.6%	25.5%	14.7%	0.6%	0.6%
(*B*) Type covering your own knowledge with greatest justice	24.8%	29.3%	26.6%	14.7%	4.5%
(*C*) Type ranking the class most accurately as to knowledge	25.5%	38.8%	23.6%	9.6%	2.5%
(*D*) Type which encourages most rote memory	14.6%	3.8%	32.5%	44.6%	4.5%
(*E*) Type giving greatest confidence that review and study masters materials which will be emphasized	32.5%	19.7%	32.5%	7.0%	8.3%
(*F*) Type furnishing greatest incentive to careful evaluation of materials covered	29.9%	11.5%	26.1%	19.1%	13.4%
(*G*) Type preferred if considerable part of grade depends upon final examination	34.4%	42.7%	15.3%	5.7%	1.9%

Brinkley's findings. Brinkley summarized his study of the preference of pupils for various types of examinations in the following table.[1]

TABLE 21

PUPIL PREFERENCES IN REGARD TO TYPE OF EXAMINATION

OLD TYPE VS. NEW TYPE	NO. OF PUPILS	PER CENT
Essay Only	22	16
Essay and New Type Combined	67	51
New Type Only	43	33
Total	132	100
CHOICE OF NEW-TYPE	**NO. OF PUPILS**	**PER CENT**
True-false	20	20
Multiple-choice	71	70
Completion	1	1
Word or Phrase Answer	5	5
Arrangement	4	4
Total	101	100

Other studies. Several minor studies are brought together here as a tabulation:

AUTHOR	NO. OF STUDENTS	RATIOS OF PREFERENCE FOR NEW-TYPE TO OLD-TYPE
Bardy*	242	95: 5
Kinder†	200+	97: 3
Kolstoe‡	300	89:11**
May††	260	(Only two objected to new-type.)

*J. Bardy, "An Investigation of the Written Examination as a Measure of Achievement with Particular Reference to General Science" (1923), University of Pennsylvania.

†J. S. Kinder, "Supplementing Our Examinations," *Education*, Vol. XLV (1925), pp. 557-566.

‡S. O. Kolstoe, "Reactions to True-False Tests," *School of Education Record*, University of North Dakota, Vol. XI (1926), pp. 54-55.

**About one-ninth preferred essay-type, and about one-fourth opposed the true-false.

††M. A. May, "Measuring Achievement in Elementary Psychology and in Other College Subjects," *School and Society*, Vol. XVII (1923), pp. 472-476 and 556-560.

[1]S. G. Brinkley, *Values of New-Type Examinations in the High School* (New York, Columbia University, *Teachers College Contributions to Education*, No. 161, 1924), p. 100.

Summary. The following comments will summarize the studies abstracted in this chapter:

1. Somers has shown the relative unpopularity of examinations in comparison with other collegiate activities.
2. Gallagher's work seems to verify Somers's general findings.
3. Teachers and students seem to be agreed that examinations are relatively unpleasant tasks.
4. Somers found that, as a result of experience, the new-type tests were at least slightly more favorably received than were the traditional examinations.
5. Somers found that teachers and students differed widely in their views on the functions of examinations. Gallagher's results were similar.
6. Hughes and Buckner found positive evidence that students think the objective examination superior to the essay-type.
7. Brinkley's results indicate that a combination of old- and new-type examinations is the most desirable practice. This seems to be a reasonable conclusion.
8. Bardy, Kinder, Kolstoe, and May find that preference for new vs. old types runs in the ratio of about nine to one.
9. There are sufficient disagreements in the matter of pupils' attitudes to warrant further study.
10. If tests and examinations are unpopular, does this not indicate the need for an attempt to integrate such measurements into the complete plan of instruction? Perhaps examinations could be made less irksome by making them more diagnostic and helpful to the student. It appears that examinations might be made to offer greater incentives.

CHAPTER VI

RELATIVE VALUES OF STANDARDIZED AND NON-STANDARDIZED TESTS

Terms and definitions. Our vocabularies of educational measurement have not yet reached the point where all authors use terms with the same meanings. Such a common expression as a *standard* or *standardized* test has little meaning. As the name implies, such a test is provided with norms or standards of achievement.

If the provision for norms constitutes the sole difference between the informal (unstandardized) objective examination (new-type test) and the so-called standard test, then the latter is nothing more than an *objective or semi-objective examination with norms.* Many well-known standard tests are in fact fairly described as more-or-less objective examinations with norms.

But a genuine standard test must meet far more stringent requirements than mere possession of norms. In truth, an otherwise well-constructed standard test, with no norms at all, would fulfill most of the functions of a standard test. Following the practice of most competent standard test-makers, a standard test will be defined for present purposes as one which:

1. Has demonstrated validity resting upon some more secure basis than personal opinion. We have already listed (Chapter II) the principal methods of validating tests. *It is further assumed that each individual item has been validated separately.*

2. Has demonstrated reliability. The principal means of assuring reliability to a test have already been described

(Chapter II). We should be able to assume that a sufficient experimental selection and elimination of "dead timber" has been carried out to guarantee a reasonable approximation to the situation of having the reliability as high as is possible in view of the number of test items which may actually be included in the working time allowed. Allowances should be made, of course, for such factors as differences in subject-matter, relative economies and reliabilities of different types of test items, etc.

As a corollary of this point, it should be fair to insist that a standard test should not be placed upon the market if its reliability is so low that its uncritical acceptance by purchasers and users will result in gross mis-measurement of pupils. Unfortunately there are not a few standard tests, held rather generally in high repute, which yield reliabilites far below those ordinarily to be obtained by a thirty- to fifty-minute informal objective classroom test of the unstandardized variety. The use of such a test is probably a sheer waste of time and money in spite of the availability of norms. (See Tables 22 and 23, pages 143 and 144.)

3. Has a reasonable degree of objectivity of scoring, in order that subjectivity will not react upon the reliability and consequently the validity of the test.

4. Has norms or standards for evaluating the results obtained by the test. This requirement is less essential than the preceding ones, and much less essential than popular opinion suggests. Norms are uncertain quantities when we consider the enormous differences likely to be found in the same school subject in such situations as: rural vs. city schools, differing courses of study, differing textbooks, differing methods of instruction, differing mental abilities of pupils, etc. Any norm is at best a *mythical entity* like the "average man," the "typical pupil" etc.

Dr. A. S. Otis has drawn up a suggestive rating scale for evaluating standard tests, which helps to define our concept

of what constitutes the genuine standard test.[1] This is given on the next page. In the opinion of the present writer Otis's scale places too little weight upon validity and reliability in comparison with administrative convenience.

For a further discussion of the criteria of a good standard test, the reader is referred to Ruch and Stoddard, *Tests and Measurements in High School Instruction*, especially pages 45 to 68 and 301 to 375.

How valid and reliable are most standard tests? It is very difficult to evaluate standard tests with respect to their general validity. In some cases their validity is much higher than the classroom teacher can hope to attain with informal objective tests. In general, however, the validity of most standard tests is open to discussion when we consider that local conditions vary so greatly. Against the fact that the standard test must fit a wide variety of local situations, we can set the fact that a well-made standard test represents a much more highly refined product than is possible with the test constructed without elaborate experimentation by a classroom teacher. It is probably fair to assert that a minority of the best standard tests more than compensate for their lack of perfect conformity to local conditions, but that the rank-and-file may be viewed with considerable suspicion. It is the conviction of the author that those schools employing rather extensive standard testing programs should seek to supplement their educational measurements by locally constructed objective or new-type tests. Such schools form the minority; the larger number may well afford to adopt a systematic program of testing by standardized measures.

Considerably more important than an extension of standard testing in the United States is a critical re-exam-

[1] *Test Service Bulletin*, No. 13, "Scale for Rating Tests," (Yonkers-on-Hudson: World Book Company, 1926).

SCALE FOR RATING TESTS	NAMES OF TESTS				
Manual (5)					
Validity (15)					
Reliability (10)					
Reputation (5)					
Ease of Administration (Total 15) (*a*) Preparation (4)					
(*b*) Time limits (4)					
(*c*) Explanation needed (3)					
(d) Alternative forms (4)					
Ease of Scoring (Total 15) (*a*) Objectivity (10)					
(*b*) Time required (3)					
(*c*) Simplicity (2)					
Ease of Interpretation (Total 15) (*a*) Norms (5)					
(*b*) Directions for interpreting (4)					
(*c*) Class record (1)					
(*d*) Application of results (5)					
Convenient Packages (5)					
Typography and Makeup (5)					
Test Service (10)					
Total (100)					

ination of the measures employed at present. There are a great many tests of excellent popular repute that have been outgrown by the progress in educational measurement, and these should be abandoned in favor of more recent and more highly perfected measures. It is, of course, manifestly unwise to refer to such tests by name although recent writers on educational measurement have published experimental evidence sufficient for the critical selection of adequate testing materials.[1]

Recent books, such as those cited in the footnote, have shown the courage to present the advantages and limitations of specific tests in a critical fashion based upon actual use and statistical study.

Monroe was the first to bring together any considerable amount of data on the reliabilities of well-known tests.[2] Table 22 gives his results, the average and median being inserted by the present author. The twenty-one well-known tests show an average reliability of between 0.67 and 0.75, depending upon whether the average or median is chosen. Such a central tendency is surely disappointing. It should be pointed out that Monroe's list is far from a random sampling of tests, his list covering chiefly reading and arithmetic tests (many of them constructed some years ago), and the selections largely confined to publications of a single company. Moreover, reading tests are very difficult of construction. Also, with a few notable exceptions (standing near the top of the list), these tests are very short ones requiring from five to fifteen minutes time.

The author has gathered from his files 149 of his computations of reliability coefficients for standard tests (Table 23).

---

[1]T. L. Kelley, *Interpretation of Educational Measurements* (Yonkers-on-Hudson: The World Book Co., 1927), especially pages 214-348.
P. M. Symonds, *Measurement in Secondary Education* (New York: The Macmillan Co., 1927).
G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction* (Yonkers-on-Hudson: The World Book Co., 1927).

[2]W. S. Monroe, J. C. DeVoss and F. J. Kelly, *Educational Tests and Measurements* (Rev. ed., Boston: Houghton Mifflin Co., 1924), p. 42.

Except for about a dozen determinations the tests are high-school and junior high-school tests, the latter being confined principally to the fields of geography, history, reading, and arithmetic. Such tests are also used in grades two to six in many instances. Table 23 presents these findings in a series of columns which segregate the tests into working times of varying lengths.

TABLE 22

RELIABILITY COEFFICIENTS OF STANDARDIZED EDUCATIONAL TESTS

TEST	COEFFICIENT
Illinois Intelligence General Intelligence Scale[1]	.92
Courtis Standard Research Tests, Series B[3]	.87
Brown Silent Reading Test—Rate	.86
Courtis Silent Reading Test No. 2—Rate	.85
Otis Group Intelligence Scale[4]	.84
Monroe Standardized Silent Reading Test, Revised[1]—Rate	.84
Courtis Silent Reading Test No. 2—Comprehension—No. Quest.	.80
Starch Silent Reading Test—Comprehension—Words	.77
Monroe General Survey Scale in Arithmetic[1]	.76
Monroe Standardized Silent Reading Test Revised[1]—Comprehension	.76
Monroe Standardized Silent Reading Test Revised[1]—Rate	.75
Monroe Standardized Silent Reading Test Revised[1]—Comprehension	.72
Starch Silent Reading Test—Comprehension—Ideas	.72
Indiana Attainment Scale No. 1[3]	.66
Starch Silent Reading Test—Rate	.62
Pressey Primer Scale[2]	.59
Courtis Silent Reading Test No. 2—Comprehension—Index	.58
Pressey First Grade Vocabulary Scale[3]	.37
Brown Silent Reading Test—Comprehension—Quantity	.36
Pressey Primer Scale[2]	.33
Brown Silent Reading Test—Comprehension—Quality	.19

Average = .67
Median = .75

[1]Walter S. Monroe, *The Illinois Examination*, p. 47. University of Illinois Bulletin, Vol. 19, No. 9, Bureau of Educational Research Bulletin, No. 6. (Urbana: University of Illinois, 1921.)

[2]L. W. Pressey, "A Group Scale of Intelligence for Use in the First Three Grades: Its Validity and Reliability," *Journal of Educational Research*, (April, 1920) Vol. I, pp. 285-94.

[3]Unpublished data of the Bureau of Educational Research, University of Illinois.

[4]S. S. Colvin, "Some Recent Results Obtained from the Otis Group Intelligence Scale," *Journal of Educational Research* (January, 1921), Vol. III, pp. 1-12.

TABLE 23

RELATION BETWEEN RELIABILITY OF STANDARD TESTS AND THEIR WORKING-TIME LIMITS

(These data principally appear in Ruch and Stoddard, *Tests and Measurements in High School Instruction.*)

r	WORKING TIME LIMITS (In minutes)									TOTAL	AVERAGES OF ROWS
	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 119	120 – 179	180 – 239		
.95 – .99					2		2	1		5	83.5
.90 – .94		1	7		4				1	13	44.1
.85 – .89		1	6	3	5	2				17	35.1
.80 – .84		3	4	6	7		1			21	35.7
.75 – .79	2	2	3	4						11	22.7
.70 – .74	4	4	2	3		1				14	20.2
.65 – .69	4	4	2	1						11	14.5
.60 – .64	5	1	2	3						11	15.4
.55 – .59	4	3		4	1					12	20.3
.50 – .54	6	2	3	3		1				15	19.2
.45 – .49	3			1		1				5	20.5
.40 – .44	2	2								4	9.5
.35 – .39	3			1						4	12.0
.30 – .34	3									3	4.5
.25 – .29			1							1	24.5
.20 – .24	2									2	4.5
Total	38	23	30	29	19	5	3	1	1	149	
Averages of Columns	.55	.68	.77	.69	.86	.69	.92	.97	.92	.694	

It is apparent at once that the short tests (less than 30 or 40 minutes) are low in reliability as a rule, although there are many exceptions. The general correspondence with Monroe's table is striking. Unlike Monroe's table, Table 23 includes a number of determinations of the reliability of a long "battery" of tests (the *Stanford Achievement Test*). If results from this (somewhat more than two-hour test) were excluded, the agreement of averages would have been almost perfect.

The results of Monroe and the author show rather conclusively that *short standard tests, with some exceptions, defeat one of their principal purposes, viz., reliable measurement.*

Standard and informal objective tests compared and contrasted. The foregoing tables and discussion must not be interpreted as a disavowal of standard tests and testing. On the contrary, the author is a firm believer in the values of the standard test. If present remarks are construed as ultra-critical, and there is danger of this happening, the defense is that we have been making a plea for continued, and extended, use of standard tests which have been selected *critically*. There is no denying that the impersonal, national and normative nature of the standard test gives it a unique position in educational practices. It is unavoidable that the standard test can never hope to parallel *all* educational conditions until such time as, if ever, there shall be reasonable uniformity of practice. In the long run we shall come to know more and more definitely what elements in our curricula prepare for adult activities, pleasures, and outlooks. To the extent that this aim comes to be realized, to that extent and to that extent only can highly valid standard tests be constructed. Reliability of measurement, on the other hand, waits upon no such final determinations. We can secure reasonable approaches to accurate measurement at the present time by means of the application of the current techniques of test construction. One thing is certain, viz., that we face a decision between continuing to use the five- to fifteen-minute test, with its resulting and certain limited reliability, and the adoption of more time-consuming measures, standardized or informal. We must abandon the thoroughly untenable position that time spent in testing is time wasted in teaching. Teaching and testing are aspects of the same process. It is further beside the point to claim that standard testing is too expensive. There has never been a case in the history of education where worth-while practices have been in the long run viewed, accepted, or rejected upon the sole basis of cost. In fact, the most inexpensive practices have invariably proved to be the most costly. School budgets always prove sufficiently elastic to

cover any costs which can be demonstrated to be profitable outlays of money.

But, after all has been said and done, there are indisputable limitations of the standard tests in the complete measurement program. They will always present some degree of alienation from the local educational situation. To this extent they will require supplementing on the part of the classroom teacher. The only solution of this matter seems to rest in the locally constructed test, objective or otherwise. We are not yet ready to abandon completely the traditional examination. It has its undeniable place. We must, however, recognize its limitations, and continue also to point out the shortcomings of informal and standardized objective tests. It is doubtful whether standard tests can be made sufficiently detailed to provide constant diagnostic guidance to teacher and pupil when we consider the economic and commercial limitations imposed upon such measures. The immediate solution of our problem of educational measurement seems to lie in the combined and complementary use of standardized and locally-derived testing materials.

Part II

HOW TO CONSTRUCT AN OBJECTIVE EXAMINATION

CHAPTER VII

THE BUILDING OF AN OBJECTIVE TEST OR EXAMINATION

Analysis of the job. The general order of operations in constructing an objective test may be listed as follows:

I. Drawing up a *Table of Specifications*
II. Drafting the items in preliminary form
III. Deciding upon the scope (length)
IV. Editing and selecting the final items
V. Rating the items for difficulty
VI. Breaking the items into alternative forms
VII. Rearranging the items in order of difficulty
VIII. Preparing the instructions for the test
IX. Making the answer keys or stencils
X. Deciding upon rules for scoring

Before entering upon a detailed discussion of these ten steps or operations in building a test or examination, some general justification is needed for certain of the steps included and for their order of appearance.

It will be noted that the decision as to the scope and length of the test is made the third step rather than the first, as is commonly the case. At first sight this seems illogical. However, the length (number of items) needed in a test cannot be decided, a priori. It is only after the preliminary items have been written and the available number of good items has been ascertained, that it is possible to decide how many items are demanded for an adequate sampling of the subject-matter. It might be thought in advance that fifty items would cover a certain group of topics, but after the task of writing the items had been

finished, it might be apparent that fifty items were too few to cover the ground thoroughly or that fifty worth-while items could not be constructed. For this reason it is recommended that the decision as to the exact length of the test be postponed until the items have been drafted in preliminary form.

The sixth step is not absolutely essential, but it is a desirable extension of usual practice. The values of alternative forms for tests have been mentioned before. The advantages of duplicate forms for stabilizing grading purposes will be described in a later chapter.

I. DRAWING UP A TABLE OF SPECIFICATIONS

The term "Table of Specifications" was adopted for the sake of emphasizing the need for a general guide or skeleton in building a test. Such a table guards against the omission of essential items, the over-emphasis of minor topics, and improper balance of the sampling. The drawing-up of a working plan before drafting specific items goes a considerable distance in establishing the validity of the final test when completed.

The various steps in constructing an objective test may be introduced by an actual example.

As a more or less typical example, a six-weeks' history test over the period previous to the Revolution was chosen. The materials covered during the six weeks represented largely the first six chapters, pp. 1-125, in Beard and Bagley, *The History of the American People.* The Table of Specifications which was drawn up is shown on the next page.

Several points should be noted about this specimen table. The major topics have been numbered with Roman numerals. Each major topic is also given a key letter which stands as an abbreviation of the full topical statement. The key

TABLE OF SPECIFICATIONS: THE PRE-REVOLUTIONARY PERIOD

No.	*Topic*	*Key Letter*	*Percentages of Items*
I.	Early trade routes and commerce	T	10%
	(*a*) Transfer of center of commerce from the Mediterranean to the Atlantic		
	(*b*) The trade with the Orient		
	(*c*) Marco Polo and his influence		
	(*d*) The early navigators		
	(*e*) The problem of a water route to the East		
	(*f*) Aid of various monarchs to exploration		
	(*g*) Approximate dates		
II.	Famous navigators and explorers	N	10%
	(*a*) Columbus; life and ideas		
	(*b*) Spain's aid to Columbus		
	(*c*) The voyages of Columbus		
	(*d*) Vasco da Gama		
	(*e*) Amerigo Vespucci		
	(*f*) Magellan		
	(*g*) Cortes and Mexico		
	(*h*) Conquest of Peru by Pizarro		
	(*i*) Ponce de Leon and De Soto		
	(*j*) The French explorers		
	(*k*) Cabot, Drake, and the English explorers		
	(*l*) The Spanish Armada		
III.	European conditions which led to the desire to explore and colonize	E	15%
	. . . . .		
IV.	The colonization of America	C	30%
	. . . . .		
V.	The struggle of European nations for supremacy in North America	S	20%
	. . . . .		
VI.	Life in colonial America	L	15%
	. . . . .		
	TOTAL		100%

letters are to be used for identification in sorting the items after they have been placed on cards (to be described later). Each major topic is followed by a stated percentage. For example, Topic I (key letter, *T*) is to contribute *approximately* 10 per cent of the total number of items in the final test. These percentages are not assigned arbitrarily, but represent the teacher's careful judgment as to the proportionate values of the several major topics to be covered. The percentages are left as such, and no attempt is to be made at this time to change these into actual numbers of items. This change can be made more intelligently under the third step in the analysis at the opening of the chapter. The Table of Specifications has not been completed, as it was thought that enough detail was given to define the procedure.

It should be noted that the sub-topics are lettered (*a*), (*b*), (*c*), etc. If desired, the keying of test items can be carried still further; e. g., items dealing with the "transfer of center of commerce from the Mediterranean to the Atlantic" might be keyed as *T* (*a*). No attempt has been made to assign percentages to sub-topics; these are to serve as reminders. To carry very far the assignment of percentages would defeat its own purposes by resulting in an impracticable and inflexible scheme which could not be followed. The sub-topics can be used in thinking about the worth of test items by asking oneself such questions as: "How many good items can I make for this subtopic?" "If I can ask but one question on this, what one thing is most important?" "How does this compare with that in importance?" etc.

The *particular scheme* outlined above is not presented as the best possible, but it has been used many times by the author and his students, both in informal and standard objective-test construction. It is recommended that *some such table* be drawn up as a part of the validation of any important test to be constructed.

II. DRAFTING THE ITEMS IN PRELIMINARY FORM

With the Table of Specifications at hand, the next step is that of writing down tentative test items. In doing this, little attention need be paid to the percentages. Take each topic and sub-topic in turn and write out items which cover the "high points" of each. Do not spend much time refining the wording. The important tasks just now are:

1. Covering the field thoroughly but at the same time avoiding trivial points; and
2. Deciding which objective *technique* or *type* (true-false, completion, multiple-choice, matching, etc.) is best suited to handling the particular question in mind.

In the end it is far more economical of time and labor to place each preliminary or tentative test item on a small card rather than to write these consecutively on ordinary sheets of paper. Cards may then be rearranged, shuffled, discarded, inserted, etc., *without necessitating any rewriting of other items.* For this purpose 3x5 library cards are best; ruled if pen or pencil is used, unruled if the items are typewritten. These cards should each contain:

1. The key letter (to designate the topic)
2. The test item (double spaced to allow for corrections)
3. The indicated answer
4. A temporary sequential number. (It is convenient to have this follow the key letter.)

The samples below and on the next page are satisfactory:

T-1

The discovery of America was part of a mighty historic movement which transferred the naval and commercial power from the (Mediterranean) Sea to the (Atlantic) Ocean.

N-26

John Cabot landed on the shores of

Florida Virginia West Indies Labrador

E-11

After 1534, the Established Church of England embraced the Catholic religion.

True False

The preliminary phrasing of the test items should be done with reasonable care, although the main attention should be given to deciding the type of test technique which would be most satisfactory. The refining of phraseology can be done more economically at a later time when the final numbers of items needed have been decided upon.

Pages 156-159 give a series of preliminary test items covering the first two main topics (*T* and *N*) of our Table of Specifications. They are written in sequence, but it should be remembered that each would appear on a 3×5 card according to the recommendations of this volume.

One important rule might be laid down at this time: *In framing preliminary test items, try to make up from 25 to 50 per cent more items than your estimate indicates will be kept in the final test.* This has two advantages:

1. A great deal of "culling out" is then possible.
2. An excess of items gives greater latitude in balancing the emphasis on the major topics and in making equivalent duplicate forms, if desired.

There is one very important aspect of the test at this stage of construction, viz., the choice of the *types* or test *techniques* to be employed. Teachers commonly ask, "What is the best type to use, the true-false, the multiple-response, the completion, or some other?" This question is very difficult of answer. There are important differences between the several principal test types; these were listed in a preceding chapter. These differences include matters of reliability, susceptibility to chance and guessing, adaptability to the subject-matter at hand, economy and objectivity of scoring, etc. Some of the more moot questions will receive a rather thorough discussion in Part III (especially Chapters XI and XII) of this volume. For the present we can do little more than to make a few general comments:

1. The type of test item (technique or mechanical form) should be decided principally upon the adaptability of that technique to the particular bit of subject-matter. It will be noted that the items which are given later in illustration of the building of an objective test in history are written down in mixed form: some true-false, some simple recall, some completion, some multiple-choice, and an occasional matching exercise. The decision as to which type to use in a given case is largely a matter of judgment and experience. Certain bits of subject-matter seem to fit themselves into one of the types; others do not lend themselves very readily to any of the types. In most cases, however, a combination of two or three of the common test types will handle everything which it is essential to include in the test.

2. It is ordinarily unwise to leave the items in the scrambled arrangement of the sixty-odd preliminary items shown on pages 156-159. It is preferable that all true-false items be assembled as one part or division of the test; the same being done for each of the other kinds of items employed. Thus, in a test employing true-false, completion, and match-

ing types, it would be best to divide the total test into three divisions or parts, one for each type of item, rather than to scramble all three techniques into one undivided test. Three sets of directions are needed in either case, and it is more systematic and less confusing to segregate the different types of items, each type having its individual directions. Such segregation need not and, for reasons of economy, should not be done at this stage of test construction.

For further guidance in handling the construction of the test at this stage, the reader is referred to the reviews of experimental studies given in Chapters XI and XII, Part III.

The preliminary draft of potential test items follows.

T. Early Trade Routes and Commerce

T- 1. The discovery of America was part of a mighty historic movement which transferred naval and commercial power from the (Mediterranean) Sea to the (Atlantic) Ocean.

T- 2. The Mediterranean Sea may be regarded as the "cradle" in which the great nations of the ancient world were "born." True False

T- 3. The last of the great ancient nations was Greece. True False

T- 4. During the Middle Ages (Italy) was the country which was the chief center of trade.

T- 5. The interest in commercial things prevented Italy from developing much art. True False

T- 6. Italy connected the markets of China and India with those of Paris and London. True False

T- 7. During the Middle Ages the language of educated persons was (Latin).

T- 8. The power of ancient Rome and Greece gradually shifted to Spain, Portugal, France, and England. True False

T- 9. The earliest Spanish and Portuguese navigators were chiefly interested in finding lands across the Atlantic which could be colonized. True False

T-10. The early navigators and merchants were seeking new (trade routes) to (China) and (India).

T-11. Rome fell in (476) A. D.

T-12. The Crusades were pilgrimages to the Holy Land. True False

T-13. The Crusades greatly stimulated trading with the countries of Asia. True False

T-14. Marco Polo was a famous monk. True False

T-15. Marco Polo lived for many years in ___(China)___. When he returned to Europe, he told many tales about the vast ___(riches)___ of the Orient.

T-16. Marco Polo was one of the leaders of the Crusades. True False

T-17. Little was known about the Far East before the birth of Columbus. True False

T-18. The Italian geographers of 1450 A. D. thought that Asia could be reached by sailing around the southern point of ___(Africa)___.

T-19. The invention of the ___(compass)___ was a great aid to navigation.

T-20. Water routes to the Orient were found earlier than were good land routes. True False

T-21. The land routes to India and China were unsatisfactory due to the slowness of travel, the danger from attack by ___(robbers)___, and the high tributes demanded by the ___(rulers)___ of the lands through which the merchants passed.

T-22. When Rome "fell," the Roman Empire was invaded by (barbarous) tribes from northern ___(Europe)___.

T-23. Spain was at one time conquered by the Moors of Northern Africa. True False

T-24. The kings of the former provinces of the Roman Empire, although often tyrannical, were less oppressive than the feudal lords. True False

T-25. The kings of the various European countries did little to encourage the development of navigation and commerce. True False

T-26. At the time of the discovery of America, Italy was a nation in name only, being in reality a collection of small city and state governments. True False

T-27. It is believed that a Norseman by the name of Eric the Red first discovered America about 1000 A. D. True False

T-28. Prince Henry of Portugal was a famous ___(navigator)___.

T-29. The explorations of the Atlantic Ocean begun by the Italians were carried on by the ___(Portuguese)___.

T-30. The southern point of Africa is called the Cape of ___(Good Hope)___. The first to sail around this cape was ___(Diaz)___, a Portuguese seaman.

T-31. Arrange the following events in order of their occurrence. Mark the first one *1*, and the next *2*, etc.

( ) Landing of the Pilgrims
( ) Plundering of Rome by the barbarians
( ) Discovery of America by Columbus
( ) Voyage of Eric, the Red
( ) The Crusades

N. Famous Navigators and Explorers

N- 1. Columbus was born in Madrid Naples Genoa Florence
N- 2. Japan was also known by the name of Palos Zipango San Salvador The Azores
N- 3. A famous poem celebrating the voyage of Columbus was written by the American poet, (Joaquin Miller).
N- 4. The first land sighted by Columbus was one of the (Bahama) Islands.
N -5. Columbus sailed under the flag of Portugal. True False
N- 6. The long-sought water route to India was first found in 1497 by da Gama Columbus Diaz Balboa
N- 7. The name "America" comes from the name of an Italian sea captain, (Amerigo) (Vespucci).
N- 8. Columbus died in ignorance of the fact that he had discovered a new world. True False
N- 9. Pinzon was the first white man to see the Pacific Ocean. True False
N-10. (Magellan) was the first to reach the Pacific Ocean directly by sailing across the Atlantic.
N-11. Magellan was killed in a fight with the natives of the Philippine Islands. True False
N-12. The Strait at the southern end of South America was named for Columbus Balboa da Gama Magellan
N-13. A Spaniard by the name of (Cortes) discovered the country of Mexico.
N-14. The first country discovered by the Spaniards which really had the much sought-for riches was (Mexico).
N-15. Cortes treatment of Mexico may be described as kindly advisory robbery co-operative
N-16. The early missionaries found that the Mexicans were adherents to the Catholic religion. True False
N-17. The ruler of Mexico at the time of the visit of Cortes was (Montezuma).
N-18. Cortes found a very low form of civilization when he visited Mexico. True False
N-19. The conquest of Peru was led by Cortes Pizarro De Soto Ponce de Leon
N-20. Pizarro's treatment was very much like that of Cortes for Mexico. True False
N-21. Ponce de Leon and De Soto were less fortunate in finding riches in Florida than were Pizarro and Cortes in South America. True False

N-22. De Soto finally reached the __(Mississippi)__ River where he ___(died)___.

N-23. The Southwest was explored first by De Soto Coronado Cortes Verrazano

N-24. The exploration of the St. Lawrence River and surrounding territory was first undertaken by ___(Cartier)___ and __(Champlain)__.

N-25. The last of the great nations of Europe to join the exploration of the New World was England France Spain Portugal

N-26. John Cabot landed on the shores of Florida Virginia West Indies Labrador

N-27. Sir Francis Drake may be described as a coward pirate statesman colonizer

N-28. The Spanish fleet was known as the ___(Armada)___.

N-29. The breakdown of Spain's rule of the sea was accomplished by a great ___(naval)___ defeat by the ships of __(England)___.

N-30. Number the following events 1, 2, 3, etc., in the order in which they occurred.

( ) Conquest of Spain and Peru
( ) Defeat of the Spanish Armada
( ) Columbus' first voyage to America
( ) Marco Polo's travels
( ) Invention of the mariner's compass

III. DECIDING UPON THE LENGTH OF THE TEST

Judging roughly from the number of items yielded by the first two chapters of Beard and Bagley (about thirty pages), it appeared that it was quite feasible to make at least 250 preliminary items on the period prior to the Revolutionary War. Allowing a shrinkage of fifty items (more or less), it is estimated that 200 suitable items might be secured, if needed. These 200 would exhaust the subject rather thoroughly. It would thus be possible to make a test of 200 items, or *two forms* of the same test with 100 items each.

The making of two forms is to be preferred over a single longer form for these reasons:

1. If it is thought advisable to give 200 items, both forms can be administered.

2. It is almost certain that a few pupils will be absent when the test is given. To use a second form as a "make-

up" (if one is available) will be fairer to all, provided the two forms are almost exactly equal in difficulty.

3. The second form can be used for re-tests on pupils who wish to "make up" their low grades on the regular test.

4. The two forms may be used in rotation, year after year, and thus provide a basis for comparing successive classes without serious danger of coaching or cramming effects.

The discussion will assume from now on that *two* forms of the test will be made, each form having 100 items. It should be noted that this decision as to the length of the test (in terms of numbers of items) was made *after* information was at hand as to such facts as (*a*) approximate number of worth-while items which could be made, and (*b*) the numbers needed to cover the subject thoroughly.

IV. EDITING AND SELECTING THE FINAL ITEMS

This is the "culling out" stage of the test. It is performed, preferably, a day or two after the preliminary drafts of the items were made, in order to edit and revise these rough statements with a fresh and critical mind. This editorial stage is by far the most critical step in the construction of the test, unless we except the second operation (drawing up the preliminary items and selecting the test type to use).

The teacher must scrutinize each test item much as an editor criticizes every line of an important manuscript. The test-maker should put himself, so far as possible, in the attitude of the pupil. Try to misread the meaning to see if there are possible misinterpretations that would mislead or prove ambiguous. Keep in mind that *good sentence structure* is a prime requisite for a valid test item. See if an easier synonym can be found for any difficult word or term. See that the punctuation is such as to assist in making the intent of the test item clear. In types of tests like the completion or simple-recall, write down every answer that you can

think of as likely to be given by a pupil. List those for which you will give full credit, likewise those for which no credit will be given. (Avoid giving half-credits.)

There is little that can be said in words which will guide the test-maker at this stage. Experience is the only safe guide in the long run.

It may help to clarify procedures to study in some detail a few of the preliminary drafts of items as given on pages 156-159 for the projected history test.

Item T-1. This item seems to be unambiguous. It probably should be retained as it summarizes what might be thought of as one of the great world movements of all history. It is more than a fact question; it expresses a broad concept of the sweep of historic events.

Item T-2. This is similar to the preceding item. It calls for thought, as neither the wording nor the fact is stated in even similar form in the text. Keep this one.

Item T-3. This is more nearly a fact question. It is, however, an important question likely to catch the pupil who is "cloudy" on ancient history. A child might know that Greece and Rome were both great nations but be unaware that Rome was the conqueror of Greece and the successor of the latter as the ruler of the world.

Item T-4. This seems to be a valid item. It is largely a matter of fact, but it bears importantly upon the movements which led to the discovery of America.

Item T-5. Less important, perhaps. Might go out if an excess of items is found to be the case. It is phrased to catch the not-too-alert child.

Item T-6. Should be discarded or revised. The word "connected" is too abstract a phraseology for elementary school pupils. Dull pupils might think of the matter as a physical connection. It might be improved as follows: "Italian shipping connected . . . etc."

Item T-7. This is fact—but surely an important one.

Item T-8. Another item similar to T-1 and T-2, but of considerable importance in the broad outlines of history.

Item T-9. Pupils often confuse the earlier period of the search for riches with the later colonization movement. This item will help to show up such a confusion.

Item T-10. A good item except that the first blank (trade routes) will elicit numerous doubtful responses. The idea of trade routes seems to be a unitary idea, and the acceptance or rejection of responses other than the one marked as acceptable will not be difficult in most cases.

Note that in such items, *either* order for "China" and "India" is acceptable. This is a general rule covering all such cases. "America" should be marked wrong. "Japan" is acceptable.

Item T-11. This is not important enough to demand the *exact* year (on the part of grammar-school pupils). Discard, or change to "Rome fell between 400 A. D. and 500 A. D.," or some less exact statement.

Item T-12. This should be valid.

Item T-14. Probably should not be used. It is far-fetched. It looks like a premeditated attempt to "fool" the pupil. It would be less objectionable in the form: "Marco Polo was a famous monarch monk traveler general."

Item T-15. The first blank is satisfactory. The second blank may elicit some responses which will be hard to score. Keep or discard according to final needs.

Item T-19. Certain other nautical instruments may occasionally be mentioned, but these should receive full credit.

Item T-21. Not of the greatest importance, but a factor in stimulating the search for less hazardous routes. "Pirates" will be given instead of "robbers," but such a response is worthy of credit. Likewise "princes," "kings," etc., will be given on the second blank. These should receive full credit.

Item T-23. Unimportant for present purposes. Discard.

Item T-24. Not important for purposes of understanding the backgrounds of American history; better reserved for the course in ancient history in the high school.

Item T-26. Similar to T-23 and T-24.

Item T-28. Unimportant. A better form, perhaps, would be: "Prince Henry of Portugal was known as the 'Navigator.'" True False

This change, however, does not increase its importance.

Item T-31. Such exercises are valuable if the chronological decisions are not too close. The present items are not objectionably close together in historical sequence. If two forms of a test are to be made, several more such matching tests should be prepared so that there are at least two or three such exercises in each form of the test. Otherwise it will be wasteful to write directions for and to administer a single five-item matching test.

The foregoing comments may or may not help the reader. They do typify, to some extent at least, the frame of mind which the test-maker must assume in criticizing his work.

Since it was decided to make two forms of this test with 100 items per form, it is necessary to eliminate eleven items if we follow the percentage allowance of the Table of Specifications. (I. e., 10% of 200 = 20.) The author will not attempt to do this. It is sufficient to note that there is somewhat more than a fifty per cent excess of items falling under this main topic. This should guarantee reasonably well that the twenty finally selected are important and valid.

V. RATING THE ITEMS FOR DIFFICULTY

The advantage of having the items of a test in increasing order of difficulty was discussed in Chapter II. The procedure for such ratings is simple, but difficult in the sense that such ratings are not very accurate since they are at best highly subjective estimates.

The rating may be done on either a 5- or 10-point scale, the former probably being fully as accurate as the latter. Whichever scale is used, the procedure is as follows:

1. Rate "1" those items which are so easy that all or nearly all of the pupils may be expected to answer them correctly.

2. Rate "5" (or "10" depending upon the scale used) those items which you think will be failed by all or nearly all of the pupils.

3. Assign the intermediate ratings ("2" to "4," or "2" to "9") to those intermediate in difficulty, so far as you are able to distribute the ratings by approximately equal intervals.

4. Write the ratings on the item cards. (Note: if two or more teachers co-operate in rating the items, be sure to use the same scale. In this case, place the ratings on the back of the card, each teacher being assigned a corner for her rating and being cautioned to make her rating before turning over the card and thus exposing to view any previously recorded ratings.)

It is assumed that all eliminations to the desired numbers have been made before the ratings are carried out. The rating may be done first, however; this procedure has the advantage of making possible the elimination at once of any excess number of items rated as too easy or too difficult.

VI. BREAKING THE ITEMS INTO EQUIVALENT FORMS

The history test (used here as an illustration) may be constructed in two roughly equivalent forms as follows:

1. Throw out all doubtful, too-easy, too-difficult, or otherwise unsatisfactory items until the numbers shown by the Table of Specifications are approximated. (This is really Step IV in our outline, and if these eliminations have already been made, there is nothing further to be done about selecting the items.)

2. Deal the items into two (or more, as the choice may be) piles exactly as playing cards would be dealt. The intention here is to equalize the forms through the law of *chance*. The final items are still in topical arrangement before the process of breaking into duplicate forms is begun. Therefore, if the items of the first topic are thus subdivided, then the next topic, etc., each form will receive approximately equal numbers (equal samplings) of each major topic (such as were indicated by the key letters *T*, *N*, *E*, etc., in our illustration).

Another procedure which accomplishes the same purpose would be as follows:

1. After the eliminations to the final numbers have been made, renumber the items consecutively, beginning with the first item of the first main topic and proceeding through the other main topics in order.

2. Throw the odd-numbered items (Nos. 1, 3, 5, 7, etc.) into one form and the even-numbered items (Nos. 2, 4, 6, 8, etc.) into the second form. If three forms are to be made, the procedure is obviously slightly different, thus:

Form A	*Form B*	*Form C*
Item 1	Item 2	Item 3
Item 4	Item 5	Item 6
Etc.[1]		

[1]It should be noted carefully that we are considering the case where items are arranged topically *and not in order of increasing difficulty*. The breaking into equivalent forms is a different matter in the latter case. If, as is often the case in standard test construction, the items are arranged in order of difficulty *before* they are broken into forms, the procedure would be:

For Making 2 Forms		*For Making 3 Forms*		
FORM A	FORM B	FORM A	FORM B	FORM C
Item 1	Item 2	Item 1	Item 2	Item 3
Item 4	Item 3	Item 6	Item 5	Item 4
Item 5	Item 6	Item 7	Item 8	Item 9
Item 8	Item 7	Item 12	Item 11	Item 10
Etc.		Etc.		

This plan should be studied carefully, as it is designed to prevent systematic differences in difficulty of the different forms.

The differences between the dealing of (1) items in chance order of difficulty and (2) those in increasing order of difficulty are exactly analogous to (a) the usual dealing of shuffled playing cards and (b) dealing in turn of a pack of cards which have been first arranged in order of *increasing* value. Under (*b*) player No. 2 would always receive a slightly better card than player No. 1, and his resulting hand would be systematically better than that of No. 1.

It should be noted that assignment by chance to duplicate forms *can only be depended upon when fairly large numbers of items are involved.* The breaking of 100 items, by chance, into two forms of fifty items each may be expected to yield forms that differ not more than from two to five points in average difficulty; occasionally the difference will be more, but more often less. When 200 items are broken into two forms of 100 each, the expected variation in difficulty in the resulting forms would be relatively less. In any case, if from 100 to 200 items are broken into two forms by chance, the resulting inequality of forms will be markedly less than if successive examinations are constructed de novo each year.

VII. REARRANGING THE ITEMS IN ORDER OF DIFFICULTY

If the items of the test (or of the individual forms) have already received difficulty ratings, it is a simple matter to rearrange them in increasing order of difficulty. It has already been pointed out that this step, if taken, will increase the reliability of the test through such means as increased motivation, a better distribution of time and effort on the part of the pupils, etc.

VIII. PREPARING THE INSTRUCTIONS FOR THE TEST

The variations in the instructions given for objective tests are legion. Most test authors have their preferred forms of stating such directions. The important things about test instructions are clarity, fullness, and brevity consistent with the first two requirements.

The amount of detail needed will depend largely upon (*a*) the familiarity of the pupils with tests of the type being given, and (*b*) the ages or mentality of the pupils. In grades two to four or five it is decidedly better to use the method of written instructions which are read silently by the

members of the class while the teacher (or other examiner) reads the instructions aloud. After the pupils have become "test wise" from repeated contacts with objective or standard tests, the instructions may be abbreviated except where new test techniques are employed.

The following general rules may help to emphasize the significant features of a good set of directions:

1. In writing instructions, phrase the directions so as to meet the level of the lowest mentalities in the group.

2. Use the simplest synonyms for all words or ideas. With very young children it may occasionally be necessary to sacrifice grammatical construction in the interest of clarity e. g., in a two-choice test it is probably permissible to instruct pupils to select the "best" answer rather than the "better," since "best" is the normal expression of young children. Critics occasionally object to calling test items (which are usually complete or incomplete declarative sentences) "questions." This criticism is beside the point with very young pupils. Moreover, in the traditional examination, as well, the word "question" has always included both imperative and interrogative sentences.

3. Be generous in the use of samples, especially with young or backward pupils. The examination is a measure of achievement, not of ability to understand and to follow directions. In the case of certain standard tests many of the zero and very low scores arise solely from inadequate instructions, the inadequacies being chiefly undue brevity and adult terminology.

4. Where the test technique is complicated, such as is the case with many matching tests, multiple-response tests with numbered alternatives, some kinds of cross-out tests, etc., use a fore-exercise or practice test to supplement the printed or verbal directions. The *National Intelligence Test* is a good example of the use of fore-exercises in a standard test.

5. The instructions should direct the pupil *where and how* to record his answers. The samples should also show the same facts. The pupils should be told whether to hurry or to work slowly and carefully. If the test is timed, the pupils should be told in advance what the time allowance is.

6. It is gradually being conceded that the instructions should inform the pupil about the answering of doubtful and unknown questions. De Graff and Ruch have found some evidence (Chapters XI and XII) that it is more valid to instruct pupils not to guess when the answering reduces to pure guessing.

In cases of doubt, but where the pupil has some "hunch" or inkling as to the probably correct answer, it seems better to allow him to follow his "hunches." The experimental evidence on this point is rather meager and authorities differ. Dr. Ben D. Wood has always used instructions against pure guessing. Dr. W. A. McCall has taken the other point of view upon the theory that the more the guessing, the more adequate the statistical correction for guessing or chance. The work of DeGraff and Ruch tends to support Wood's position rather than that of McCall, although the issue is not as important as some seem to think. Teachers have objected to encouraging guessing as a matter of bad habit formation.

The following sets of test directions are thought to be reasonably adequate in the main.

1. True-False

Below are 50 true-false statements. About one-half of them are true and about one-half are false. Read each statement carefully. If you think it is true, draw a line under "true." If you think it is false, draw a line under "false."

Take each question in order, but do not waste too much time on one that you do not know. Skip it and go on to the next. *Do not guess!*

If you have any time left you may go back and work on those you left out.

Ask no questions after the signal to "Go" is given.
Study the samples carefully before beginning actual work.
[Samples follow]

In case the words "true" and "false" do not appear after each item, it is customary to place a dotted line either at the extreme left or right of the item. The pupil is then instructed to record his judgments by some such markings as:

		If true	*If false*
(*a*)	.....	T	F
(*b*)	.....	True	False
(*c*)	.....	Yes	No
(*d*)	.....	+	0
(*e*)	.....	+	−

Of these (*a*) is not very satisfactory, for two reasons: (1) T and F look too much alike for accurate and rapid scoring, and (2) when pupils correct their own tests, it is too tempting to change the T into an F by adding a short mark.

Numbers (*d*) and (*e*) seem the best both for speed and clarity. There is little to choose between the two. High-school pupils who have studied algebra may feel that the plus and minus method is slightly more meaningful. When self-correction is allowed, (*d*) is greatly to be preferred over (*e*) since pupils can change − to + too easily during the process of correcting the papers.

2. Multiple-choice

There are five possible words given for completing each incomplete statement below. Only one of these words makes the statement true.

Read each question carefully, decide which word makes the truest completion, and then draw a line under that word, as shown in the samples.

If you do not know the answer to any question, do not waste time on it, but go on to the next. Do not hurry, as there will be enough time for all to finish.

Look at the three samples before you begin to work.
[Samples follow]

When the multiple-choice test requires responding by writing the *number* of the correct response, the wording might be:

Below are 75 incomplete statements. Five words or phrases are given after each statement. One of these five words or phrases makes the statement true; the other four are incorrect.

Read each incomplete statement carefully, decide which of the five possible words or phrases makes the truest sentence, then write the NUMBER (not the word or phrase itself) on the dotted line at the right, as shown in the samples.

Samples:

(*a*) The best temperature for living rooms is (1) 50° (2) 60° (3) 68° (4) 75° (5) 78° 3

(*b*) The blood is pumped by the (1) liver (2) lungs (3) stomach (4) veins (5) heart 5

Begin here.

3. COMPLETION

Certain words have been left out in the sentences or paragraphs given below. Dotted lines show where the words are left out. In most cases *just one word* has been left out.

You are to write, on the dotted lines, the words which have been left out. The three samples show you how to do this test.

[Samples follow]

Do the rest of the sentences or paragraphs in exactly the same way as shown in the samples.

The *Paragraph Meaning* test (a completion test) of the *Stanford Achievement Test* gives a set of directions which have proved to be adequate in high-first and low-second grades. This test uses the double method of spoken directions by the teacher and silent reading by the class.

Read the words at the top of the page, here. (Hold up booklet and point to the sample sentence.) **It says** (read slowly): **Dick and Tom were playing ball in the field. Dick was throwing the ball and . . . . .** (pause) **was trying to catch it. Who was trying to catch the ball?** (Encourage pupils to answer aloud.) As soon as correct answer is given, say: **Yes, Tom was trying to catch it. You must write *Tom* on the dotted line.** (Pause until word is written.)

Wherever you see a dotted line on these two pages, it means that a word has been left out. Begin with No. 1, read each sentence carefully, and write JUST ONE WORD on each dotted line to show what has been left out. When you have finished the first page, go right on to the second page. Ready—Go.

The pupil's test booklet (which is before the child as the teacher gives the above directions) looks like the following:

Stanf. Adv. Exam. A

TEST 1. READING: PARAGRAPH MEANING

Sample: Dick and Tom were playing ball in the field. Dick was throwing the ball and.....was trying to catch it.
Write JUST ONE WORD on each dotted line.

---

1. Fanny has a little red hen. Every day the hen goes to her nest and lays an egg for Fanny to eat. Then she makes a funny noise to tell Fanny to come and get the .....
2. A kitten can climb a tree, but a dog cannot. This is very lucky for Nellie's kitten. Every time Joe's big dog comes along the kitten climbs a tree and the .....cannot follow.
3. Etc.

4. Matching Tests

The column at the left below gives the names of ten men. The column at the right gives ten events *connected with* the names of the ten men.

Look at each event in the column at the right; then find the man in the column at the left who is connected with that event.

Place the Number (*not the name*) of the man in front of the event with which he is associated. The first one is already done correctly as a sample.

Men	Answer	Events
1. George Washington	7	Inventor of the cotton gin
2. Harvey W. Wiley	.....	Louisiana Purchase
3. Thomas Jefferson	.....	Forest conservation
4. Robert E. Lee	.....	Development of banking system
5. William J. Bryan	.....	Pure food laws
6. Alexander Hamilton	.....	President of the Confederacy
7. Eli Whitney	.....	First President of the U. S.
8. Robert Fulton	.....	Advocate of free silver
9. Gifford Pinchot	.....	Inventor of the steamboat
10. Jefferson Davis	.....	Commander of the Southern armies

In case of incomplete matching (where one column contains an excess of terms in order further to reduce chance successes), certain changes will have to be made in the above instructions.

These suggested instructions for four of the principal types of objective tests must suffice for presenting the fundamental principles in phrasing directions to the pupils. The exact phraseology is of small importance; the main issues being clarity, completeness, and the generous use of samples to supplement the written instructions. With young children, reading in unison and marking sample items under the direction of the examiner are most helpful.

IX. MAKING THE ANSWER KEYS OR STENCILS

The choice of an economical answer key or stencil depends principally upon (*a*) the nature of the test to be scored and (*b*) the number of tests to be scored.

The labor of scoring may be made very small indeed if the needs of economical scoring are kept in mind when the test is planned. Certain tests like the multiple-response may be quite laborious in scoring even when the most convenient scoring devices are provided, but may be planned so as to obviate from fifty to ninety per cent of such labor by better arrangements of the test items proper.

We need to recognize two principal types of devices for indicating responses, as follows:

1. *Aligned* response columns, usually vertical in position, e. g.,

Snowbound was written by .....
The date of the birth of Shakespeare was .....

2. *Staggered* response blanks (or positions), e. g.,

The purified blood returning to the heart from the ..... enters the ..... auricle and from there it passes through the ..... valves into the ....., etc.

One of the principal products of China is corn gold wheat tea ironwood

Ohio is bounded on the west by Missouri Indiana Iowa Illinois Michigan

Aligned responses are always more economical and are always possible with simple recall, matching, and true-false tests. They are often possible with multiple-choice tests (when the method of response is by number, not underlining). The force of these suggestions may be felt by studying the following fragments of tests.

I.

1. A little ..... is a dangerous thing.—*Pope.*
2. To me the ..... flower that blows can give
   Thoughts that do often lie too deep for tears.—*Wordsworth.*
3. Honor and shame from no condition rise;

   ....
   ..... there all the honor lies.—*Pope.*

II.

1. The formula for sulphuric acid is ..... H_2SO_4
2. Chlorine belongs to the group known as ..... Halogens
3. Alkaline solutions turn phenolphthalein ..... red

III.

1. The mosquito is known to carry typhoid malaria small-pox yellow fever
2. The most important class of foods for tissue-repair are the proteids minerals fats carbohydrates
3. The trunk is divided into two main cavities by means of the ribs œsophagus trachea diaphragm

IV.

1. The American Revolution began in (1) 1762 (2) 1775
   (3) 1783 (4) 1789 (5) 1812 2
2. Cornwallis surrendered at (1) Yorktown (2) Jamestown
   (3) Saratoga (4) Appomattox (5) Valley Forge 1
3. The President of the Confederacy was (1) Lee
   (2) "Stonewall" Jackson (3) Thomas (4) McClellan
   (5) Jefferson Davis 5

V.

1. All bacteria are injurious. True False
2. The bodies of all plants and animals are made up of cells. True False
3. The chief use of the red blood corpuscles is to kill disease germs in the blood. True False

VI.

1. Stevenson wrote *Treasure Island.* True False
2. Dickens was a writer of lyric poetry. True False
3. *She Stoops to Conquer* was written by Byron. True False

VII.

1. A straight line is the shortest distance between two points. +
2. In the expression $37y^4$, y is an exponent. −
3. If $x^2+9=0$, x equals -3. −

These seven examples are all common practices. The advantages and limitations of each call for brief comment.

No. I is not economical of scoring. It is about as good a device as can be adopted for such a test, however, It will require several times as long *per blank* as No. II, but it probably cannot be improved greatly without grave complications.

No. II is usually termed "simple-recall," and is a form of completion test. It is the most rapidly scorable type of completion test. Where conditions permit, it is to be preferred over No. I. Note, however, that teachers often place the terminal blanks *immediately* after the close of the statement. This makes the blanks occur in a staggered arrangement very unsatisfactory for scoring. The recommendation is to align the blanks in the simple-recall tests, even if the statements vary greatly in length. If desired, there can be hyphen leaders (.....) or dots (.....) inserted between the end of the statement and the response lines, thus·

The formula for sulphuric acid is..... H_2SO_4
Chlorine belongs to the group known as.....Halogens

In typewritten material the best practice is dots with a solid line for the response. If the material is to be set in type, dot leaders should be used with hyphen leaders for the response line.

No. III is to be compared with No. IV. No. III is simpler, and perhaps better adapted to children below the sixth grade because more simple instructions may be written. No. IV is by all means more desirable from the standpoint of economy of scoring. The device employed in No. IV was invented by Dr. Arthur S. Otis, and represents the most rapidly scored multiple-choice test technique ever developed. A variation is the use of (*a*), (*b*), (*c*), etc., instead of (1), (2), (3), etc., for labelling the responses. Letters have the advantage in tests in mathematics or history information where numbers (dates) occur frequently. This plan of numbered (or lettered) responses occasionally leads to error when the pupil selects the right response but writes the *wrong* number (or letter). For this reason it is chiefly adapted to high-school and college levels.

Nos. V, VI, and VII should be compared. No. V is markedly inferior to the other two. No. V has no advantage except the very slight one (Cf. Nos. III and IV in this respect) that, in the case of Nos. VI and VII, there is some danger of the eye failing to "carry" out to the correct blank or response word. The answer is consequently misplaced one or more blanks either up or down. Scorers should be on their guard against this. No. VII is faster than No. VI, but the difference is not great enough in many circumstances to make it a vital factor. Note that + and 0 are often used instead of + and − to indicate true and false, respectively. Sometimes the method of responding is by writing "T" and "F" on the blanks. This is probably inferior to plus and minus. Both the + and − and the T and F (written in) methods are not well adapted to having pupils correct their own papers. It is too easy to change a − into a + or a T

into an F when correcting the papers. The use of + and 0 is somewhat better for self-correction; although underlining is also reasonably satisfactory. Many test workers prefer to have the +, −, or 0 placed between the item number (at the left) and the first word of the statement.

A few minutes study of the samples of test techniques given in Chapters VIII and IX will reveal which ones are best adapted to rapid, accurate, and easy scoring.

We can now classify answer keys and stencils as follows:

1. Strip keys for aligned vertical columns of response blanks or response words in such tests as:

(*a*) Simple recall
(*b*) Numbered multiple-choice
(*c*) Matching
(*d*) True-false (especially the + −, the +0, or the writing of T and F, etc.)

2. Transparent celluloid or tissue-paper stencils for such tests as:

(*a*) Unnumbered staggered multiple-response
(*b*) True-false, yes-no, same-opposite, etc., when underlined

3. Cut-out stencils for such tests as:

(*a*) Staggered (ordinary) completion
(*b*) Staggered computation

4. Answer sheets for reference (not to be superimposed or aligned directly on the test sheets). These are sometimes used for almost any variety of tests.

1. The Strip Stencil. Fig. 8 shows a strip stencil applied to a page of test material of the numbered multiple-response type. The stencil is merely a strip of heavy paper or cardboard from one-half to one inch wide and the length of the test page. The easiest way of making such a strip stencil

Section 1

ENGLISH, LITERATURE, ANI

ANSWER KEY
Page 1

1. Snowbound was written by (1) F (3) Whittier (4) Tennyson (5) Kiplin	3	3	
2. The Gettysburg Address was given by (2) Daniel Webster (3) General Grant (5) Thomas Jefferson	1	1	
3. The Pied Piper ridded Hamelin of (3) rats (4) frogs (5) beggars	3	2	—
4. The god who held up the heavens was cules (3) Odysseus (4) Mercury (5	1	1	
5. The best known work of Coleridge (2) Ancient Mariner (3) Ode to a Grecia (5) The Excursion	2	2	
6. A phrase "a government of the people, the people" was uttered by (1) Washingtor houn (4) Lincoln (5) Wilson	4	5	—
7. Gulliver's Travels is the story of (1 America (2) one of the first African expl a missionary (4) the struggles of a ma (5) the imaginary adventures of an English s	5	5	
8. Robinson Crusoe is noted for (1) meaning (2) its careful presentation of sc clear and life-like story (4) its political intimate revelation of the hidden life of the	3	3	
9. The Wreck of the Hesperus was wr (2) Longfellow (3) Riley (4) Stevenson	2	4	—
10. The House of Seven Gables was writ (2) Hawthorne (3) Poe (4) Cooper	2	2	
11. The Pit and the Pendulum was wri (2) Whittier (3) Holmes (4) Bryant	5	5	
12. One of Robin Hood's men was (1) I (3) Little John (4) Bill Sykes (5) Mi	3		—
13. The Call of the Wild was written by ling (3) London (4) Stevenson (5)	3	3	
14. The ferryman of the Styx was (1) Cl (3) Argus (4) Scylla (5) Typhon	1	4	—

FIG. 8.—A strip stencil for scoring an English test.

is to write down the correct answers on an extra test-sheet. Then place the strip immediately to the left of the column of answers and write each answer on the strip. The illustration shows the stencil applied to an actual paper, the marks in the right hand margin representing errors or omissions.

The scoring of a paper with such a stencil reduces to a process of comparing paired numbers, letters, etc. The ordinary objective examination, if open to such scoring, can be scored in from one to two minutes, depending upon its length, the number of different pages, and other factors. The author once had occasion to score several thousand fifteen-page booklets, of which Fig. 8 shows the first page. There were 400 items in the test, all of the type shown. The better scorers exceeded the rate of two items per second, making it possible to check 400 items on fifteen pages, count the number of correct answers, and record the score on the title page in less than five minutes per booklet after a little experience. This rate is many times more rapid than that of the reading of any essay examination of comparable scope.

Fig. 9 shows three other strip stencils. Stencil A was planned for use with a non-staggered simple-recall test. Stencil B is used with a plus-zero response, true-false test. Stencil C is used with a true-false test where T and F are encircled or underlined. The adaptability of the strip stencil to other aligned forms of tests needs no further comment.

2. The transparent stencil. Transparent stencils may be made either of tissue paper or of celluloid sheets such as were formerly much used in side curtains of automobiles. The celluloid is much to be preferred, although it is somewhat expensive. Where large numbers of tests of the staggered type are to be scored, the celluloid stencil is the only very feasible device. The million or two of intelligence tests given in the United States Army during the World War were scored by means of such stencils.

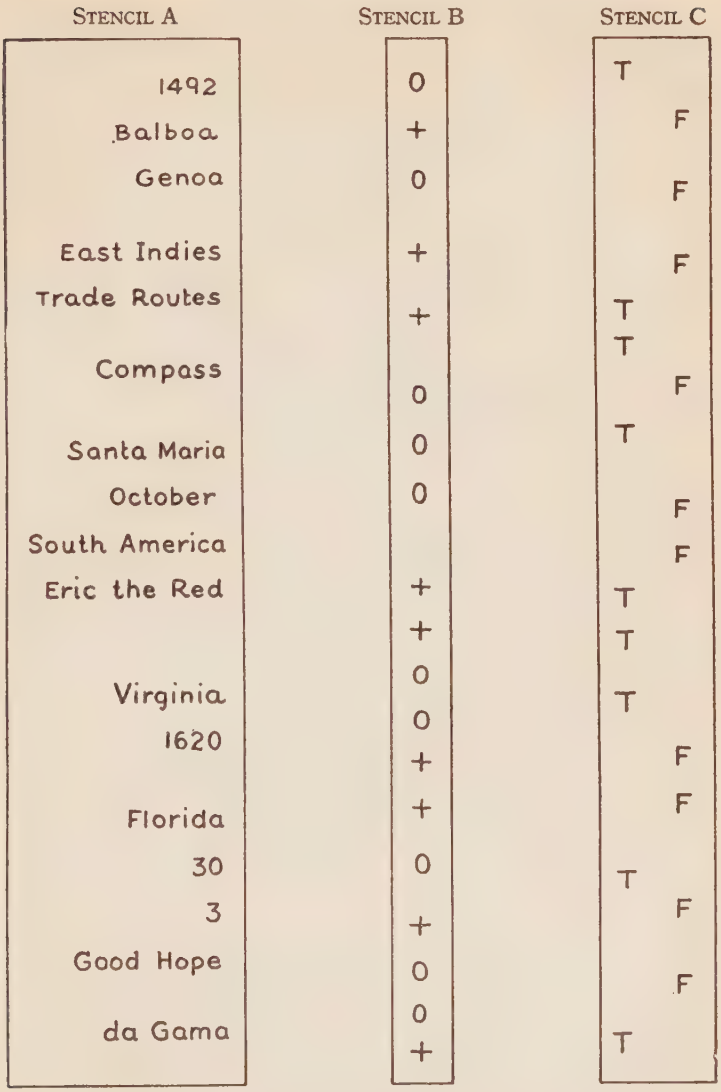


FIG. 9.—Various types of strip stencils.

In making a transparent stencil the first step is to mark an extra set of the test papers or booklets so that the correct answers are underlined (or written in, etc., as the case may be). Then place the celluloid or tissue paper sheet directly over the test page thus marked. Assuming that the responses are indicated by underlining, place dots on the transparent stencil in such a position that these dots fall directly on the middle points of each underscoring on the test page below. Care must be taken to keep the stencil in exact place while making the dots. Launderer's ink is best for such stencils. After the dots are dry, dip the stencil in white shellac and hang it up by one corner for a few minutes to dry. Such a stencil is almost indestructible and may be used thousands and thousands of times.

Fig. 10 shows how such a transparent stencil will look if superimposed upon an actual test page; the printed or mimeographed test material is not shown, although of course it does show through such a stencil. Wherever a dot (on stencil) appears to bisect the line (on pupil's paper), the answer is correct. If a dot and line fail to superimpose, the answer is wrong. If a dot appears but no line, the item was omitted. Note how quickly the *four* errors (three errors and one omission) on this one page can be detected. With a little practice, the eye will cover such a stencil-test assembly very rapidly, and with slight danger of serious error. The practiced scorer using transparent stencils soon becomes almost an automaton, the failure of dot and line to superimpose actually seeming to "strike him in the face" as the eye travels down the page.

Such stencils may be made of tissue paper or waxed paper, a good quality of waxed paper having the advantage that it is fairly transparent while still having considerable strength and resistance to folding and tearing. Such paper stencils have short life and are less convenient to handle than the more firm and durable celluloid sheets.

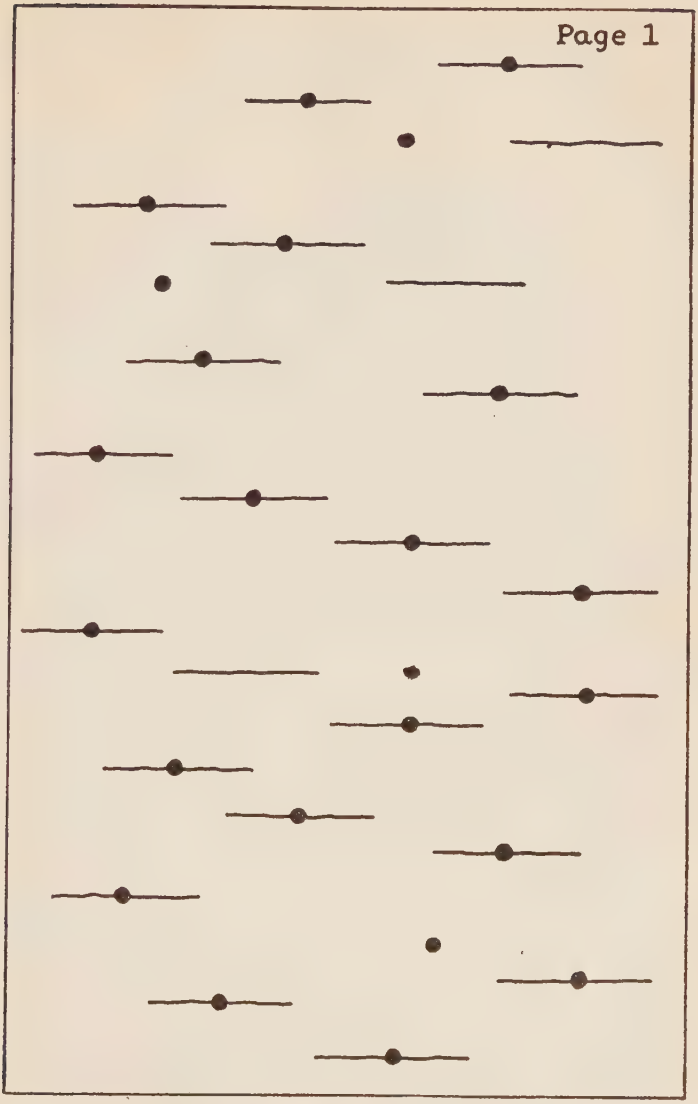


FIG. 10.—A transparent celluloid stencil.

Teachers who follow our suggestions for duplicate examinations to be administered in alternate years can well afford to make and keep on file scoring stencils.

3. Cut-out stencils. These will serve much the same purposes as the transparent stencil. They are somewhat more trouble to prepare, but are less expensive than the celluloid stencils. As before, the first step is to write in the correct answers on an actual test-sheet or series of sheets, particularly in the case of completion tests, computations, etc. Then take a sheet of thin cardboard and a piece of carbon paper. Place the marked test-sheet upon the cardboard (the same size as the test sheet) with the carbon sheet in between. Draw rectangles around each answer, making the rectangles large enough to include the space ordinarily required by a pupil's answer. Remove the cardboard sheet and cut out the rectangles as indicated by the lines drawn. Below the opening of each rectangle, write the answer which would appear in the opening above the rectangle in case the pupil answered the item correctly. Fig. 11 shows how such a cut-out stencil would look when superimposed upon an actual test page.

Note that two errors appear on this test page. These have been checked in the right-hand margin.

Such stencils may be used for a wide variety of tests, in general, for any staggered arrangement of test responses. Arithmetic or algebraic computation tests lend themselves to such scoring.

4. Answer sheets for reference. Under average conditions the teacher who has to score but twenty to forty papers at a time will often feel that she need not prepare an elaborate stencil like those already described. For short tests the answers can be memorized quickly as the result of scoring a dozen or so tests. In such cases a list of answers for reference will suffice. In the case of completion tests

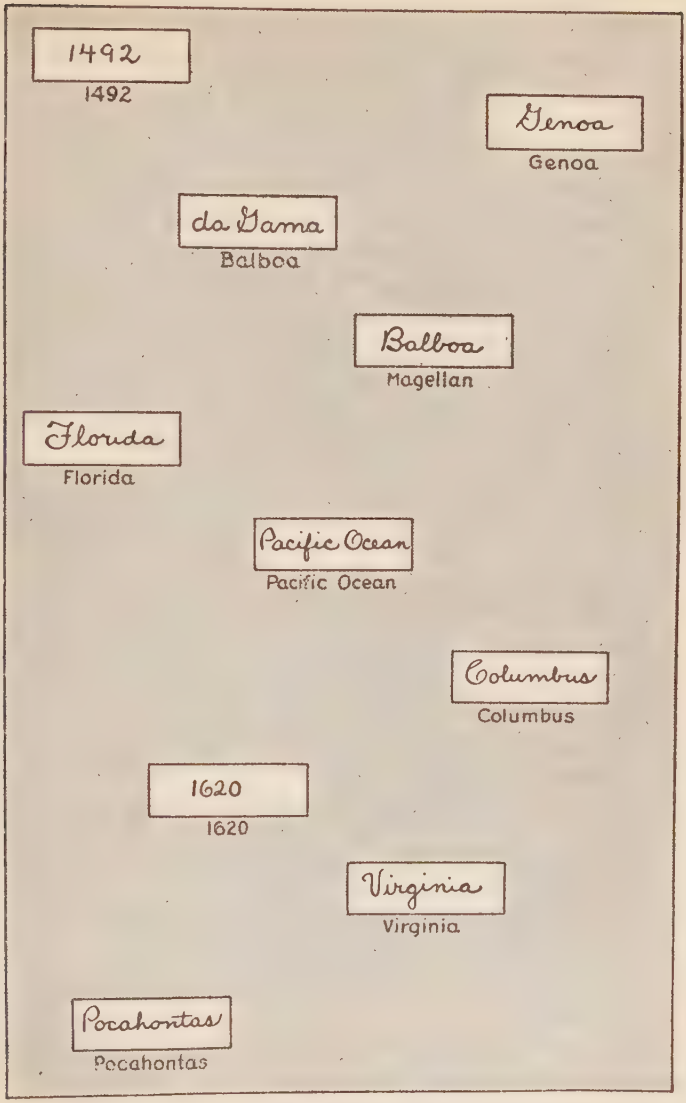


FIG. 11.—A cut-out stencil for scoring a history test.

it is almost necessary to keep a list of answers which have actually been found among the pupils' replies. Two such lists should be jotted down: (*a*) those answers for which credit is granted, and (*b*) those answers deemed unworthy of merit. Such lists must be *growing* lists, since each additional paper is likely to raise new issues as to scoring.

It is ordinarily unwise to use any plan of scoring which calls for actual reading of the test items and the pupils' responses. In objective tests it is sufficient to scrutinize *only* the response; any scoring plan which requires attention to the printed or mimeographed test item proper is certain to be wasteful. Some completion tests require actual scrutiny of pupils' responses, especially in case the answers are unusual and do not appear on the stencil.

A reference sheet may often be placed in close enough juxtaposition to the actual test sheets to allow for fairly rapid comparison of recorded and acceptable answers.

X. DECIDING UPON RULES FOR SCORING

Certain aspects of scoring rules have already been touched upon in connection with the preceding sections. There is little more to add at this time, although Part III of this volume re-introduces certain controversial issues related to scoring of objective tests.

A few general statements must suffice for the present.

1. Avoid giving *partial credits*. Except in rare cases, mark an answer either right or wrong.

2. Give each test item *one* point of credit in such tests as true-false, simple recall, and multiple-choice.

3. In completion tests give one point for each blank which is correctly filled.

4. In matching tests give one point for each pair correctly matched.

5. Do not attempt to weight test items for difficulty or relative importance. Such weightings are quite as likely to result in injustice as in justice.

6. Where *chance alone* gives the pupil the opportunity of getting one-half or one-third of the items correct by pure guessing, apply the chance correction formula. In two-response tests (including the true-false, yes-no, same-opposite, etc.) and in three-response (multiple-choice) tests, the scores should be corrected for chance.

The formula for correcting for chance effects is:

$$\text{Score} = \text{No. Right} - \frac{\text{No. Wrong}}{(n-1)},$$

or more simply: $S = R - \frac{W}{(n-1)}$, where:

S is the corrected score

R is the number of right (correct) responses

W is the number of wrong responses

n is the number of possible responses presented to the pupil for each item.

In two-response tests (including the varieties of the true-false), this formula[1] becomes: $S = R - W$

For three-response tests, this formula is: $S = R - \frac{1}{2}W$

Similar formulas may be derived for four, five, or more responses, but in practice, no correction for chance is ordinarily employed when *more than three* responses are presented, i. e., for four-, five-, or more, response tests. It should also be noted that with tests of mixed character, (i. e., when certain items present two alternatives, certain present three choices, and certain others four or more choices), there can be no method of making allowance for guessing and chance effects.

[1]The formula $S = A - 2W$, where A represents the number of items attempted, is algebraically equivalent to $S = R - W$, and is usually somewhat more convenient to use.

The use of the correction formula for true-false (and other two-response tests) is illustrated below.

Method I

150	(Possible or maximum score)
13	(Omissions)
137	(Attempts, i. e., total number of items attempted)
26	(Wrongs, i. e., total number of items answered incorrectly)
111	(Rights, i. e., total number of items answered correctly)
26	(Wrongs)
85	(Score, i. e., rights minus the wrongs)

Method II

It should be noted that the formula $A-2W$ (attempts minus two times the number wrong) gives the same result, thus:

$137-2(26)=137-52=85$ (Score)

In explaining to a class the method of scoring two-response tests, the former (and longer) procedure is less likely to be misunderstood than is the latter, which *appears to* take off two for each wrong answer. The second procedure is likely to be regarded as a double penalizing by pupils.

Since it is usually desirable to return papers to the class, at least for inspection, pupils are certain to raise the question of how the score was found. It is worth while to explain the method in detail to pupils of high-school age, although it is probably wasted effort to attempt to justify the logic of the $R-W$ scoring to elementary-school pupils. When the issue is raised, try to make the pupils see the reasonableness of the right-minus-wrong method of scoring from the standpoint of probability. Chapters XI and XII should be read in this connection.

In calculating corrected scores on two- and three-response tests, it is necessary to count but two of the three possible

responses, viz., the rights, the wrongs, and the omissions. In two- or three-response tests of about the "ideal"[1] degree of difficulty, the number of rights is likely to exceed both the wrongs and omissions. It is somewhat easier, therefore, to count only the wrongs and omissions.

The method of correcting three-response tests for chance is illustrated by the following example:

125	(Possible or maximum score)
9	(Omissions)
116	(Attempts)
21	(Wrongs)
95	(Rights)
$10\frac{1}{2}$	(21 ÷ 2 or one-half of the wrongs)
$84\frac{1}{2}$	(Score)

The fraction is usually dropped.

Four-, five-, or more, response tests are not corrected for chance in actual practice.

[1] "Ideal" difficulty is here defined as meaning that the average pupil will earn a *corrected* score which is roughly half of the maximum score.

CHAPTER VIII

ILLUSTRATIVE TYPES OF OBJECTIVE TESTS

Classification of types. Conneau, working under the direction of Rice and Ruch, analyzed in detail 375 objective or new-type examinations submitted in competition for prizes in a national contest for constructing such tests.[1]

Table 24 gives a brief summary of the tendencies in objective test construction. This is the only extensive summarization of practice which has appeared to date. The conditions (a contest for cash prizes) under which these examinations were constructed guarantee a standard of excellence which is undoubtedly higher than the average objective classroom test.

Table 24 shows clearly that five types of objective tests (completion, true-false, multiple-response, matching, and identification exercises, and their variates) make up over ninety per cent of the 45,418 test items included in 375 typical examinations. This does not mean that other forms are unimportant. Each school subject has its peculiar needs, and individual types of tests may occur frequently in one subject and never appear in certain others. Examples of this are to be found in mathematics, where computations and problems lead all other types, although such items were seventh in the total lists. Other cases are map location in history and geography, translation and scansion in languages, reproduction of poems, laws, and axioms in English, science, and mathematics, etc.

---

[1]A. Conneau: *Tendencies in Objective Testing in High-School Subjects as Shown by Analysis of a Representative Sampling of Such Tests* (1928), Unpublished M. A. Thesis, University of California.

The best of the 375 tests, together with certain statistical data, are published in G. M. Ruch and G. A. Rice, *Specimen Objective Examinations* (Chicago: Scott, Foresman and Co., 1929). This reference will give valuable suggestions to the teacher who wishes to compare her own tests with those which have been adjudged as having high merit by competent critics.

TABLE 24

AN ANALYSIS OF THE PRINCIPAL TYPES OF OBJECTIVE TESTS IN ACTUAL USE THROUGHOUT THE UNITED STATES (After Conneau)

TYPES OF TESTS	ACTUAL NUMBERS	PER CENTS
1. Completion tests, and variates	13,492	29.71
2. True-false, and variates	10,956	24.12
3. Multiple-choice, and variates	7,473	16.45
4. Matching, association, etc	4,845	10.67
5. Identifications, with or without diagrams	4,165	9.17
6. Correct form, including re-writing for grammar, capitalization, etc	1,254	2.76
7. Computations and problems	805	1.77
8. Rearrangements, mixed sentences, etc	803	1.77
9. Translations, in foreign languages	347	0.76
10. Reproduction from memory, poems, axioms, etc	347	0.76
11. Essay questions, short paragraphs	274	0.60
12. Map locations	264	0.58
13. Analogies	201	0.44
14. Constructions, with figures or diagrams	74	0.16
15. Deductions, of conclusions or principles from stated premises	55	0.12
16. Redundancies or cross-outs	45	0.10
17. Pronunciation, scansion, etc	18	0.04
Totals	45,418	99.98
No. of examinations analyzed 375		

As a working classification, the following outline of test types is given, the classification being based on the principal usages found by analysis of the 375 examinations:

I. Recall types
   (*A*) Simple-recall
   (*B*) Completion
   (*C*) Short-answer
II. True-false types
   (*A*) True-false (also: + − and +0)
   (*B*) Yes-no
   (*C*) Right-wrong
   (*D*) True-false-doesn't-say (also: true-false-doesn't know and true-false-can't-tell, etc.)

(*E*) Converse true-false (in mathematics only)
(*F*) True-false with diagrams
(*G*) Synonym-antonym

III. Multiple-response (or multiple-choice) types
(*A*) Multiple-response, proper (subdivided according to the number of responses presented, for example: two-response, three-response, four-response, etc.)
(*B*) Best-answer

IV. Matching exercises
(*A*) Perfect pairing
(*B*) Imperfect pairing
(*C*) Multiple matching

V. Analogies

VI. Rearrangement types
(*A*) Chronologies
(*B*) Order of operation
(*C*) Mixed sentences

VII. Computations
(*A*) Examples
(*B*) Problems

VIII. Constructions
(*A*) Mathematical figures
(*B*) Diagrams (as in science)

IX. Identifications
(*A*) With drawings
(*B*) Without drawings

X. Reproductions from memory (poems, axioms, laws, formulas, symbols, equations, etc.)

XI. Correction of errors (in grammar, punctuation, capitalization, spelling, etc.)

XII. Redundancies or cross-outs

XIII. Map location

XIV. Deduction of conclusions from premises

XV. Translations

XVI. Miscellaneous and mixed types

ILLUSTRATIONS OF VARIOUS TYPES

The following pages give a number of test fragments illustrating the preceding sixteen types of exercises. These samples are numbered consecutively for convenience of reference. The types are designated by the same numbers and letters as were used in the foregoing classification. These samples are unedited. Occasional mechanical changes have been made in the interests of varying the forms. These test fragments are to be viewed as illustrations of *techniques* rather than content. The content, however, is in every case some teacher's judgment of worthwhile material.

I. RECALL TYPES

IA. Simple Recall

Sample 1. (*Geography*)

After each state in this list, write its largest city:

Indiana	.....	Illinois	.....
Washington	.....	Colorado	.....
Missouri	.....	Iowa	.....
Oregon	.....	Alabama	.....

Sample 2. (*Business arithmetic*)

1. Telephone lines with more than one phone are called ..... lines.
2. An ..... is an itemized statement of goods sold, and is sent when the goods are shipped.
3. The one on whom a draft is drawn is called the .....

Sample 3. (*English*)

1. Women first acted in plays during the ..... century.
2. ..... wrote novels in letter form.
3. ..... burlesques the typical 18th and 19th century heroine.

Sample 4. (*Literature*)

1. *Snowbound* was written by .....
2. The god who held up the heavens was .....
3. The best known work of Coleridge is .....
4. "The shrub was like a sheeted specter" is an example of .....
5. *Pilgrim's Progress* is an example of the type of writing known as .....

Sample 5. (*Poetry*)

1. "God's in his heaven:
   All's right with the .....!"—*Browning*
2. "I, the ..... of all the ages
   In the foremost files of time."—*Tennyson*
3. "....., rest! thy warfare o'er
   Sleep the sleep that knows not breaking."—*Scott*

Sample 6. (*Spanish*)

Write on the dotted line the form of the present tense which corresponds to the subject pronoun.

1. comprar yo .....
2. hacer él .....
3. abrir nostros .....
4. Etc.

Sample 7. (*Mechanical Drawing*)

1. A drawing of a complete machine is called a(n) ..... drawing.
2. In machine drawing, perspective has ..... value.
3. To put threads in a hole is called .....
4. Small screws designated by number instead of diameter are called .....

I*B*. Completion

Sample 8. (*Physiology*)

The human body is composed of small divisions known as ....., although ordinarily we cannot ..... these small divisions without the aid of a .....

Sample 9. (*Typewriting*)

1. The spacing from the letterhead to the date ranges from ..... to ..... spaces, depending on the ..... of the letter. The spacing from the date line to the inside address is the ..... as from the ..... to the ....., or a ..... of ..... to ..... spaces.
2. The letters "i" and ..... are struck with the ..... finger of the ..... hand.

Sample 10. (*English*)

1. A tragedy is the portrayal of .... which is bound to end ..... because of .....

2. The first act of a play has three functions: ....., ..., and .....

3. At the beginning of the story the Primrose family lived in ....., where Mr. Primrose was ..... Their greatest friends were the ..... with whom they were particularly intimate, because ..... This friendship was broken up when ..... Immediately after this misfortune George Primrose went ..... Later on, like the author Goldsmith, he .....

Sample 11. (*Cooking*)

1. In making cream of tomato soup, ... may be added to the ..... to neutralize the .....

2. There are ..... tsp. in one tbsp., ... tbsp. in one ..... cup, and ..... c. in one qt.

3. Sugar digests ..... and for this reason irritates ..... and .....

Sample 12. (*Geometry*)

1. A surface has ..... dimensions; ..... and .....

2. A line has ..... dimension; .....

3. A point has ..... dimension; its one property is .....

4. Triangles are classified with regard to angles as ....., ..., and ..... triangles. They are classified with regard to the length of the sides as ....., ..., and ..... triangles.

5. If the sum of two angles is a straight angle, they are called .... angles.

Sample 13. (*History*)

A Greek ship loaded with silk from the Orient comes north along the Ionian cities of Asia Minor, through the strait of ..... to the European seaport ..... at the entrance to the Black Sea. Here the silk is traded for ....., a chief export of that region. The loaded ship returns through the strait to the ..... Sea, across this sea to Piraeus, the seaport of ..... in the district of ..... Here the cargo is again exchanged for ....., an important export of this district. Etc.

IC. Short-Answer Types[1]

Sample 14. (*English*)

Explain in one sentence what connection each of the following people had with Goldsmith:

1. Griffiths
2. Dr. Milner
3. Sir Joshua Reynolds
4. Jessamy Bride
5. Newberry
6. Contarine
7. Dr. Johnson
8. Garrick

Sample 15. (*Latin*)

Identify each of the following in a single sentence.

1. Galba 2. Pompey 3. Crassus 4. Dumnorix

Sample 16. (*Geometry*)

Write a clear brief statement giving the best reasons you know why each statement is true.

1. A diagonal divides a parallelogram into two equal triangles.
2. All equilateral triangles are similar.
3. A diameter is the greatest chord that can be drawn in a circle.
4. Every triangle may be inscribed in a circle.

II. TRUE-FALSE TYPES

IIA. True-False

Sample 17. (*Physiology*)

Underline the "true" or "false" according to your judgment of the truth of each statement.

1. Tetanus (lockjaw) germs usually enter the body through open wounds. *True* *False*
2. Pneumonia causes more deaths in the United States than tuberculosis. *True* *False*
3. White blood corpuscles are more numerous than are the red ones. *True* *False*

Sample 18. (*History*)

Draw a circle around the "T" or the "F" depending upon whether the statement is true or false.

1. Lincoln's first purpose in entering upon the Civil War was to free the slaves. T F

[1]The term "short-answer" has been used by some writers as a synonym for objective or new-type tests, etc. It is used here to mean any not-too-long answer of at least moderate objectivity of scoring.

2. Pinchot headed the food conservation program of the U. S. during the World War. T F
3. The Dred Scott decision was concerned with the question of free silver. T F

Sample 19. (*Shorthand*) To be marked + or −.

..... 1. R, L, P, and B are all downward letters.
..... 2. The circle vowel is written on the inside of curves.
..... 3. The vowels in *mean*, *she*, and *day* are joined according to the same rule.

Sample 20. (*English*) To be marked + or 0.

1. The story of Beowulf lets us know something of the Anglo-Saxon ideals. .....
2. Beowulf was without doubt written during the time of Chaucer. .....
3. Little worth-while drama was produced during the Elizabethan age. .....
4. As a whole Puritan literature lacked romantic ardor. .....
5. Satire was a prominent element in the literature of the classic period. .....

Sample 21. (*Spanish*)

1. Viva en una casa.	*True*	*False*
2. El padre de mi padre es mi abuelo.	*True*	*False*
3. Mi escuela está en Nueva York.	*True*	*False*
4. El dinero no le gusta a nadie.	*True*	*False*
5. Hace calor en el verano.	*True*	*False*

Sample 22. (*Manual Training*)

1. You can tell a rip-saw from a cross-cut saw by the size of the teeth, the rip-saw teeth being larger. .....
2. No. O sandpaper is smoother than No. 1 sandpaper. .....
3. A six-penny nail is longer than a $2\frac{1}{2}''$ screw. .....
4. Shellac is made from the dry sap of an oriental tree. .....

II*B*. Yes-No Type

Sample 23. (*Cooking*)

1. Is fish higher in protein content than beef?	YES	NO
2. Are deep-fried foods harder to digest than those fried in a small amount of fat?	YES	NO
3. Is gelatin a pure protein food?[1]	YES	NO

[1]Some teachers prefer the question form as being less likely to "fix" false notions in the minds of the pupils. (It may be questioned whether the interrogative form really helps.)

Sample 24. (*Geography*)

1. Did steam improve the transportation of the 17th century? *Yes* *No*
2. Does dairy farmland cost less than grazing land? *Yes* *No*
3. Does Oregon produce more lumber than Texas? *Yes* *No*

Sample 25. (*Latin*)

1. Is "populus Rōmānus" used in the plural? .....
2. Is "littera" used in the singular when it means an epistle? .....
3. Is "mīlia" always followed by the genitive of the things enumerated? .....
4. Is "neutrī puerī" the correct Latin translation for the phrase "neither boy"? .....

Sample 26. (*Manual Training*)

1. Spiral reamer flutes turn opposite from those of a drill. YES NO
2. The lands on a rose reamer are relieved. YES NO
3. Unequal lips on a drill cause oblong holes. YES NO
4. The web of a drill lies between the two cutting edges. YES NO
5. The flank of a tooth is below the pitch circle. YES NO

IIC. Right-Wrong Type

Sample 27. (*Sentence Structure*)

1. Glancing down the famous street, signs of every kind were visible. *Right* *Wrong*
2. Hills that have witnessed from the time of the first inhabitants all the exciting events that form the glorious past of the city and the Spanish conquests in search of gold, the craving which drew the first American settlers to this country. *Right* *Wrong*
3. Looking down from my high position, I saw, as night came stealing over the brae, a flickering of candlelight in the windows. *Right* *Wrong*

Sample 28. (*Punctuation and Capitalization*)

Check the correct ones with an "R" and the incorrect ones with a "W."

..... 1. I had never heard anyone sing "Where is My Wandering Boy Tonight?".

..... 2. With a condescending air, she handed the biggest package to Elmer, my escort, and ordered him to carry it for her.

..... 3. She pretends to be very intellectual. One of her first gestures

was to ask me if I did not think *Paradise Lost* more interesting than *Nize Baby?*

II*D*. True-False, Didn't-Say

True-False, Don't-Know

True-False, Can't-Tell

Sample 29. (*Geometry*)

Mark the following statements "T," "F," or "D" according to whether they are always true, always false, or sometimes true (doubtful).

1. Two triangles are congruent if—
   (*a*) Three sides of one triangle are respectively equal to the sides of the other triangle. T F D
   (*b*) Three angles of one triangle are respectively equal to three angles of the other triangle. T F D
   (*c*) Two sides and the included angle of one triangle are respectively equal to the two sides and the included angle of the other triangle. T F D

II*E*. Converse True-False Type

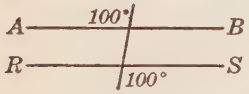
Sample 30. (*Geometry*)

If the converse of each of the following statements is true, underline the words "converse-true." If the converse is not true, underline the words "converse-false."

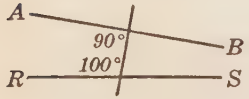
1. In the same circle or in equal circles equal chords subtend equal arcs. Converse-true Converse-false
2. If a line divides two sides of a triangle proportionately, it is parallel to the third side. Converse-true Converse-false
3. Congruent figures are necessarily equal in area. Converse-true Converse-false

II*F*. True-False With Diagrams

Sample 31. (*Geometry*)



1. Are *AB* and *RS* parallel? YES NO



2. Are *AB* and *RS* parallel? YES NO

IIG. Synonym-Antonym

Sample 32. (*Reading Vocabulary*)

If the words of a pair mean the same or nearly the same, draw a line under *same*. If they mean the opposite or nearly the opposite, draw a line under *opposite*.

1. cold-hot.....same-opposite
2. knave-villain.....same-opposite
3. recoup-recover.....same-opposite
4. plenary-complete.....same-opposite
5. adventitious-accidental.....same-opposite

III. MULTIPLE-RESPONSE (MULTIPLE-CHOICE) TYPES

IIIA. Two-Response Type

Sample 33. (*History*)

1. The first president of the Confederacy was Lee Davis
2. The turning point of the Civil War is usually taken as the battle of Bull Run Gettysburg
3. The Dred Scott Decision was concerned with slavery tariff

Sample 34. (*Language*)

1. The children (sung, sang) several songs.
2. I thought he (did, done) it.
3. Since then I have never (run, ran) away.

Sample 35. (*Language*)

1. I $\frac{\text{saw}}{\text{seen}}$ the man today.
2. The tardy bell $\frac{\text{has rang}}{\text{has rung}}$ sometime ago.
3. She sang $\frac{\text{beautiful}}{\text{beautifully}}$.

Sample 36. (*Latin*)

Strike out the incorrect form.

1. (Domī) (Domum) est.
2. Cūr pontem (fēcerīs) (fēcisses) sciō.
3. Caesar erat (dux bonus) (ducem bonum).

IIIA. Three-Response Type

Sample 37. (*Commercial geography*)

1. A natural building stone is—cement-granite-tile
2. The Columbia River is noted for—cod-sardines-salmon
3. Most automobiles are made in—Michigan-Ohio-New York

Sample 38. (*Latin*)

1. The town was not like the city.
   Oppidum (urbī, urbe, urbem) simile nōn erat.
2. Many of the Helvetians had been wounded.
   Multī (Helvētiīs, Helvētiōs, Helvētiōrum) vulnerātī erant.

III*A*. Four-Response Type

Sample 39. (*Physiology*)

1. The normal pulse rate is about 48 70 98 112
2. The trunk is divided into two main cavities by the ribs diaphragm œsophagus vertebrae
3. The absorptive action of the small intestine is greatly increased by the villi pylorus pancreas spleen

III*A*. Five-Response Type

Sample 40. (*General Science*)

1. The freezing point on the Centigrade thermometer is −273° 0° 32° 100° 212°
2. A gas which supports combustion is hydrogen nitrogen carbon dioxide oxygen carbon monoxide

Sample 41. (*History*)

Write the *number* of the correct answer on the line at the right.

1. Peter the Great was a ruler of (1) England (2) Holland (3) Russia (4) Gaul (5) Denmark .....
2. The first state to secede from the Union was (1) Virginia (2) South Carolina (3) Delaware (4) Missouri (5) North Carolina .....
3. The minimum age for a voter is (1) 18 (2) 19 (3) 20 (4) 21 (5) 25 .....

III*B*. Best-Answer Type

Sample 42. (*Biology*)

1. Leguminous plants play an important role in nature because:
   Bacteria associated with their roots return nitrogen to the soil.
   They will grow on soil too poor to support other crops.
   The economic value of the hay crop is very large.
2. The best of these definitions of photosynthesis is:
   The action of sunlight on plants.
   The process of food manufacture in green plants.
   The process by which plants give off oxygen.

Sample 43. (*Mathematics*)

1. Adjacent angles
   ..... are always equal.
   ..... always have a common side and vertex.
   ..... if added together make 90 degrees.
2. An angle of 30° is vertical to an angle
   ..... of 30° if they have a vertex in common; the left side of the one and the right side of the other form a straight line.
   ..... of 60° if the left side of one and the right side of the other, and vice versa, form a straight line.
   ..... of 30° if they have a common side and a common vertex.

IV. MATCHING EXERCISES

IV*A*. Perfect Matching

Sample 44. (*English Literature*)

AUTHORS	ANSWERS	WRITINGS
1. Oliver Goldsmith	...5....	David Copperfield
2. Jane Austen	.....	Life of Johnson
3. George Eliot	.....	Henry Esmond
4. Matthew Arnold	.....	Tam O'Shanter
5. Charles Dickens	.....	Pamela
6. Lord Byron	.....	The Prisoner of Chillon
7. Samuel Richardson	.....	Ode to the West Wind
8. Robert Burns	.....	Sohrab and Rustum
9. William M. Thackeray	.....	Locksley Hall
10. James Boswell	.....	Mill on the Floss
11. John Ruskin	.....	Lays of Ancient Rome
12. John Keats	.....	Pride and Prejudice
13. T. B. Macaulay	.....	Eve of St. Agnes
14. Alfred Tennyson	.....	The Deserted Village
15. P. B. Shelley	.....	Modern Painters

Sample 45. (*Manual Training*)

Match each style of furniture with its characteristic.

1. Mission	.....	Maple or birch; a turned job
2. Windsor	.....	Massive and plain
3. Louis XV	.....	Elaborate with delicate carvings
4. Jacobean	.....	Delicate and graceful
5. Chippendale	.....	Spiral turnings; usually finished in "antique oak"

Upon what part of a building as given in Column Two does each part of Column One rest?

1. Floor boards	..... sills
2. Rafters	..... rafters
3. Roof sheathing	..... plates
4. Joists	..... joists

IV*B*. Imperfect Matching

Sample 46. (*History*)

MEN	CHARACTERIZING PHRASE
1. Thomas H. Benton	(5) Wrote the Declaration of Independence
2. Thaddeus Stevens	( ) For 30 years a senator from Missouri
3. George B. McClellan	( ) An immigrant who worked for political reform
4. Carl Shurz	
5. Thomas Jefferson	( ) Leader of Union Army in Peninsula Campaign
6. Miles Standish	
7. De Witt Clinton	( ) Congressman demanding harsh treatment of the South
8. Charles Sumner	
9. Sir Walter Raleigh	( ) Discoverer of the New World for Spain
10. Christopher Columbus	( ) Spent a fortune to found a colony in America
11. Vasco de Balboa	
12. David Wilmot	( ) Military man of Plymouth; celebrated by Longfellow
13. Woodrow Wilson	
14. William Bradford	( ) Massachusetts senator who denounced the "Crime Against Kansas"
15. Richard Hoe	
	( ) Governor of New York—promoted the Erie Canal

IV*C*. Multiple Matching

Sample 47. (*Literature*)

Write on the lines after each character the words (from the column at the right) which best fit that character. Each word may be used more than once.

1. Gawain	(*a*)		1. fickle
	(*b*)		2. unhappy
	(*c*)		3. powerful
2. Arthur	(*a*)		4. idealistic
	(*b*)		5. untrustworthy
	(*c*)		6. of great purpose
3. Lancelot	(*a*)		7. a gossiper
	(*b*)		8. disgusted with himself
	(*c*)		9. courteous
			10. rude (Etc.)

V. ANALOGIES

Sample 48. (*Geometry*)

1. Two points : a straight line :: ..... : a plane
2. A triangular face : ..... :: a rectangular face : .....
3. Vertex : plane angle :: edge : .....

Sample 49. (*Algebra*)

1. Exponent : a number :: index : (number, coefficient, exponent)
2. x : $3x$ :: 6 : (99, 20, 27, 12, 18, 30, 6)
3. Monomial : binomial :: binomial : (monomial, binomial, trinomial)

Sample 50. (*Ancient History*)

1. The *Book of the Dead* was to the Egyptians as the ..... was to the Persians, and as the ..... was to the Hebrews, and as the ..... was to the Mohammedans.
2. Zeus was to the Greeks as ..... was to the Romans.
   Mercury was to the Romans as ..... was to the Greeks.
   Demeter was to the Greeks as ..... was to the Romans.
   Athena was to the Greeks as ..... was to the Romans.
3. Enlil was to the Babylonians as..... was to the Persians, and as ..... was to the Hebrews.

VI. REARRANGEMENT TYPES

VI*A*. CHRONOLOGIES

Sample 51. (*History*)

Arrange these "issues" according to the order of their appearance in American political history.

( ) Reduction of the surtax
( ) Free coinage of silver
( ) Internal improvements at national expense
( ) "54-40 or fight"
( ) Entering the League of Nations

Sample 52. (*English Literature*)

Re-arrange the following events from the first two books of the *Aeneid* in the order in which Vergil tells about them.

The banquet in Dido's palace
The struggle in the palace of Priam
The death of Laocoön
Venus tells Aeneas the story of Dido

The storm off the coast of Sicily
The vision of Hector appears to Aeneas
The struggle with the band of Greeks led by Androgeus

VI*B*. Order of Operations

Sample 53. (*Manual Arts*)

1. In starting an automobile, what is the order in which the following things should be done? (Number 1, 2, 3, etc.)

..... cranking or stepping on the starter
..... turning on the ignition
..... retarding spark
..... choking
..... putting in neutral

2. In glazing a window, what is the proper order for these jobs? (Number 1, 2, 3, etc.)

..... cutting glass to size
..... putting on thick putty
..... putting glass in place
..... putting on thin putty
..... cleaning out old putty
..... painting rabbet with linseed oil
..... driving in glazier points or brads

Sample 54. (*Cooking*)

The following are the steps in making muffins. Indicate by 1, 2, 3, etc., the order in which you would perform the steps.

..... bake
..... measure and sift ingredients (dry)
..... add melted fat
..... add egg
..... assemble utensils and ingredients
..... add liquids
..... place in muffin tins

VI*C*. Mixed Sentences

Sample 55. (*Latin*)

Number the words in each sentence 1, 2, 3, etc., to show the correct word-order.

1. Insulae habent multās formās.
2. Equī trahunt carrōs.
3. Rōmānī viās bonās multās muniēbant.
4. Amīsērunt in multī bellō vītam.

VII. COMPUTATIONS

VII*A*. Examples

Sample 56. (*Algebra*)

1. *Add;* $-3x$, $-5x$, and $6x$
2. What does $(3a-2a^2-1)$ plus (a^2-1) plus $(3a^2+2a)$ equal?
3. Multiply $(2x+y)$ by $(x-y)$

VII*B*. Problems

Sample 57. (*Algebra*)

1. A rectangular field is y feet wide and 40 feet long. What will represent its perimeter?

2. A rectangle is three times as long as it is wide. If each dimension is increased by 4 inches, the rectangle will be twice as long as it is wide. Find its length and width.

Sample 58. (*Chemistry*)

1. How many liters of sulphur dioxide at standard conditions will be obtained when 52 grams of sodium acid sulphate completely react with hydrochloric acid? Atomic weights:

Sodium = 23 Hydrogen = 1 Sulphur = 32 Oxygen = 16

2. A compound contains 70% iron (at. wt., 56) and 30% oxygen (at. wt., 16). Find the simplest formula.

VIII. CONSTRUCTIONS

VIII*A*. Mathematical Figures

Sample 59. (*Arithmetic*)

1. Draw a line through point A so as to form an angle of 45 degrees with line AB A——————B

2. Construct a right triangle whose sides are, respectively, $1\frac{1}{2}$, 2, and $2\frac{1}{2}$ inches.

3. Through C, draw a perpendicular to AB. A————·————B
C

VIII*B*. Science Diagrams

Sample 60. (*Botany*)

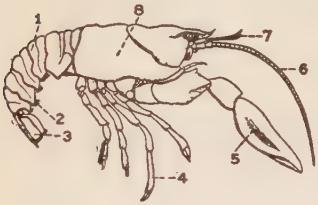
1. Draw a cross-section plan of the flower of the buttercup.
2. Draw a longitudinal section of a grain of corn.
3. Show the arrangement of the F. V. B. in the stem of a endogen.

IX. IDENTIFICATIONS

IX*A*. With Drawings

Sample 61. (*Zoölogy*)

Write the name of each structure which bears a number



1.....
2.....
3.....
4.....
5.....
6.....
7.....
8.....

IX*B*. Without Drawings

Sample 62. (*Chemistry*)

In the following list, identify by encircling E, C, M, X, respectively, the elements, compounds, mixtures, and any other classification.

1. water gas	E	C	M	X	
2. lamp black	E	C	M	X	
3. alloy	E	C	M	X	
4. ammonia	E	C	M	X	
5. air	E	C	M	X	
6. alum	E	C	M	X	
7. brass	E	C	M	X	
8. caustic soda	E	C	M	X	
9. radium	E	C	M	X	
10. steam	E	C	M	X	Etc.

Sample 63. (*German*)

Identify each of the following in a short sentence in German.

1. Brigitte
2. Die weisse Taube
3. Das Zithermädchen

Sample 64. (*English*)

Identify each of the following by naming the story in which it occurs.

1. The pillar of fire by night
2. The Sea Maid
3. The man who wanted to write a story, but was laughed at by his wife

X. REPRODUCTIONS

Sample 65. (*English*)

Quote ten lines from "As You Like It."

Sample 66. (*Chemistry*)

CHEMICAL NAME	FORMULA
1. Sodium phosphate	.....
2. Sulphuric acid	.....
3. Ammonium bromide	.....
4. Calcium carbonate	.....

XI. CORRECTION OF ERRORS

Sample 67. (*Grammar*)

Draw a line through each word which is unnecessary or incorrectly used.

1. I feel pretty good.
2. Who did you see?
3. They don't know nothing.
4. He jumped off of the car.
5. Him and I were unable to go.

Sample 68. (*Grammar*)

Draw a circle around any error or omission in spelling, punctuation, capitalization, or grammar in the following sentences. The first three are marked correctly as samples. If the sentence contains no errors, write "Correct" on the dotted line at the right. Correct each error found by rewriting the correct form on the dotted lines.

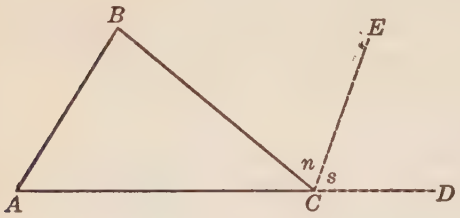
1. I saw him yesterday.	Correct
2. The meeting was called by (mr.) Jones.	Mr.
3. Who is that man(?)	?
4. Busness promises to improve.	.....
5. What a pity!, she exclaimed.	.....
6. The tardy bell has rung.	.....
7. There goes president Coolidge.	.....
8. You the leader, should go first.	.....
9. I am just wild about classical music.	.....
10. He don't appear to be very intelligent.	.....

XII. REDUNDANCIES

Sample 69. (*Geometry*)

Place, in the space to the left of each of the following statements in proof of the theorem stated, the letter "N" if the statement is a necessary step

in the proof of the theorem and the letter "U" if the statement is unnecessary to the proof. (Also correct any errors that you find.)



THEOREM: The sum of the three angles of a triangle is equal to two right angles.

GIVEN: Triangle *ABC*

TO PROVE: Angle *n* plus angle *s* plus angle *BCA* is equal to two right angles.

STATEMENTS

.....1. Produce *AC* through *C* to *D*.	.....1. A straight line may be drawn connecting any two points.
.....2. From *C* draw *CE* bisecting angle *BCD*	.....2. To draw a line parallel to a given line.
.....3. Angle *s* plus angle *n* plus angle *BCA* equals two right angles.	.....3. The sum of all the angles about a point on the same side of a straight line through that point is a right angle.
.....4. Angle *s* equals angle *B*.	.....4. If two lines are cut by a transversal, the alternate interior angles are equal.
.....5. Angle *n* equals angle *B*.	.....5. If two parallel lines are cut by a transversal, any corresponding angles are adjacent.
.....6. But angle *s* equals angle *n*.	.....6. Construction.
.....7. Therefore angle *A* equals angle *B*.	.....7. Things equal to the same thing are equal to each other.
.....8. Substitute in (3) above angle *s* for its equal angle *A* and angle *n* for its equal angle *B*.	.....8. A quantity may be substituted for its equal in any process.

A plus *B* plus *BCA* equal two right angles.

Sample 70. (*English*)

Cross out the words that make the statements incorrect.

1. Poe's literary work is remarkable for its artistic finish, realism, sadness, moral ideas, and special technique.

2. Mark Twain is best remembered for his hatred of hypocrisy, refined humor, romantic history of western life, long detailed descriptions, and strong sense of justice.

Sample 71. (*Reading comprehension*)

Cross out the word or words that spoil the sense of the sentence.

1. It was a very hot day, and I went at once into the house and put on my fur overcoat.

2. She was a beautiful girl with long curls, blue eyes, a well-shaped nose, and very even yellow teeth.

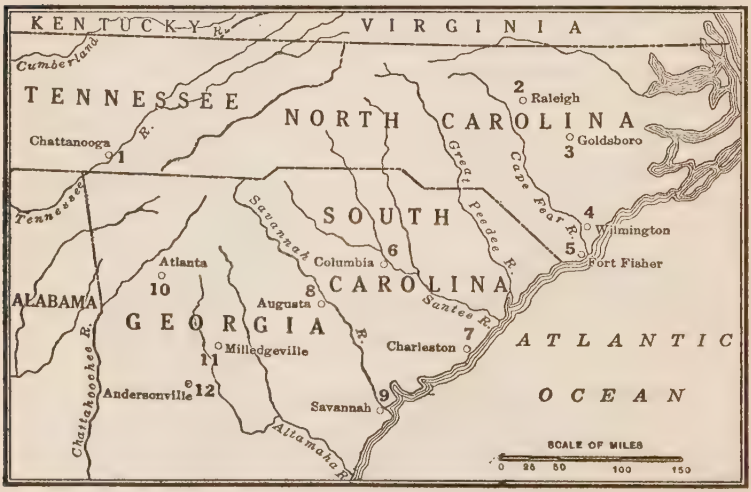
3. The man stood with his hands in his pockets as he pointed out to me the road to Boston.

XIII. MAP LOCATION

Sample 72. (*History*)

Study the map below carefully. Notice that the towns and cities shown are numbered. Write as your answer the *numbers* of the cities, in order, through which Sherman passed on his famous "March."

Your answer should include *seven* numbers.

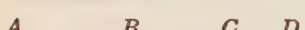


XIV. DEDUCTION OF CONCLUSIONS FROM PREMISES

Sample 73. (*Geometry*)

On the right below are a number of axioms; on the left are some figures and some statements of equality or inequality. Study the figure, the statements, and the questions. Then select the axiom from the list at the right which gives the answer to the question after each figure or statement on the left, and write within the parentheses the letter representing the axiom which correctly answers the question.

Given: $AC=BD$
Why does $AB=CD$? ( )



Given: $AB>BC$
and $BC>CD$
Why is $AB>CD$? ( )

Given: $AB>BC$
and $BC>CD$
Why is $AC>BD$? ( )

Given: $MN=OP$
Why does $MO=NP$? ( )
Etc.

(*A*) If equals are added to unequals in the same order, the sums are unequal in the same order.

(*B*) If equals are divided by equals, the quotients are equal.

(*C*) If equals are added to equals, the sums are equal.

(*D*) If unequals are subtracted from equals, the remainders are unequal in the reverse order.

(*E*) If equals are subtracted from equals, the remainders are equal.

(*F*) Halves of equals are equal.

(*G*) Like powers or like roots of equals are equal.

(*H*) If of three quantities, the first is greater than the second and the second is greater than the third, then the first is greater than the third.

XV. TRANSLATIONS

Sample 74. (*Latin*)

Write the correct Latin translation below the English sentence.

1. The woman whom you see is the queen.

.....

2. His troops will fight as bravely as possible.

.....

3. The wretched king blamed himself.

.....

Sample 75. (*German*)

Der Wolf und das Lamm

Ein Wolf und ein Lamm ſtanden am Ufer eines Baches, um zu trinken. Oben ſtand der Wolf, unten das Lamm. Kaum hatte der Wolf das Lamm geſehen, ſo fing er gleich einen Streit an. „Warum trübſt du mir das Waſſer?" ſchrie er wütend. Zitternd antwortete das Lamm: „Wie kann ich dir das Waſſer trüben? Es fließt ja von dir zu mir herab." Der Wolf ſchämte ſich; denn das Lamm hatte die Wahrheit geſprochen. Dennoch ſagte er zornig: „Vor ſieben Monaten haſt du mich geſchmäht." Sanft erwiderte das Schäfchen: „Vor ſieben Monaten war ich ja noch gar nicht geboren." „Dann hat es dein Vater getan," rief der Wolf wütend, ergriff das Lämmchen und fraß es.

Read the German story and then answer the questions in English.

1. Where did the wolf and the lamb stand? .....
2. Of what did the wolf accuse the lamb? .....
.....
3. Why was it impossible for the lamb to have done what the wolf blamed it for? .....
.....
4. What happened to the lamb at the end? .....
.....

XVI. MISCELLANEOUS AND MIXED TYPES

Sample 76. (*Punctuation*)

Punctuate the short paragraphs below.

1. Hello said the little old gentleman thats not the way to answer the door. Im wet let me in.
2. I beg pardon sir said Gluck Im very sorry but I really cant do it because my brothers would beat me to death sir. What do you want.

Sample 77. (*Appreciation of Poetry*)

DIRECTIONS: Mark the *five* (5) of the following ten passages which you feel sure would be considered as having the most beautiful language. Make your decisions in terms of your own feeling.

.....1. Her mother died when she was young
Which gave her cause to make great moan;
Her father married the worst woman
That ever lived in Christendom.

.....2. A host of Phantom listeners
That dwelt in the lone house then
Stood listening in the quiet of the moonlight
To that voice from the world of men.

.....3. A damsel with a dulcimer
In a vision once I saw;
It was an Abyssinian maid
And on her dulcimer she played
Singing of Mount Abora.

.....4. Pack clouds away and welcome day;
With night we banish sorrow;
Sweet air, blow soft, mount lark aloft,
To give my love good-morrow.

.....5. The flowers do fade and wanton fields
To wayward winter reckoning yields;
A honey tongue, a heart of gall
In fancy's spring but sorrow's fall.

. . . . .

.....10. Death, be not proud, though some have called thee
Mighty and dreadful, for thou art not so;
For those whom thou think'st thou dost overflow
Die not, poor death.

Sample 78. (*Latin*)

Identify by giving the Latin word from which the underscored word is derived.

1. Mr. Smith is the senior member of the firm. .....
2. Mary's hand was so small that she could not stretch an octave. .....
3. His vision is very bad. .....

Sample 79. (*Latin*)

Opposite each statement give the Latin construction illustrated by the word or words underlined.

1. Caesar was a man of great courage. .....
2. No one trusts those barbarians. .....
3. Labienus is in command of the legion. .....
4. The scout said that the city had been found. .....

Sample 80. (*Science*)

Cross out the word that does not belong with the others.

1. lungs breathe respiration cough swim
2. fire water burn blaze heat
3. run stand march walk skip
4. beef mutton fish pork veal
5. stomach intestines mouth ear throat

CHAPTER IX

SELECTED COMPLETE EXAMINATIONS

Introduction. Chapter VIII presented eighty short extracts representing the commonest types of objective test techniques, as found by analysis of nearly four hundred tests and examinations.[1] The present chapter will show a number of complete or semi-complete examinations selected with the following ideas in mind:

1. To show long or reasonably long tests which represent extended sampling.

2. To show a wide variety of test procedures. For this reason certain fields normally of not much interest to elementary- and high-school teachers are represented.

3. To show tests developed by classroom teachers, research agencies, official examining bodies, professional test-makers, etc.

4. To show approaches to the measurement of certain kinds of subject-matter ordinarily thought of as not being amenable to objective measurement.

5. To show methodology of test construction rather than to illustrate "ideal" content, although the content of many of the examinations represents a high level as well.

The presentation of a large number of complete examinations is very space consuming. The reader who wishes to examine other typical examinations and tests will do well to study the volume mentioned in the footnote below and certain of the references listed in the *General Bibliography.*

In some instances these examinations contain faulty items, often in violation of principles laid down in this

[1]Submitted in a national contest for twenty-five cash prizes. Thirty-five of the best of these examinations, including all the prize-winning tests, are published in G. M. Ruch and G. A. Rice, *Specimen Objective Examinations* (Chicago: Scott, Foresman and Company, 1929). This volume is planned as a companion book to the present treatment.

text. The author did not feel at liberty to make more than minor changes.

EXAMINATION I

The following series of elementary-school subject-matter tests was constructed in the office of the Superintendent of Schools of Kern County, California, under the direction of former Superintendent L. E. Chenoweth and present Superintendent H. L. Healy. Kern County has for several years prepared quarterly examinations similar to the one reproduced here for the sixth grade. Each grade has a different examination.[1]

KERN COUNTY BOARD OF EDUCATION REVIEW

1928

Let us see how well you have learned some things during the first two quarters. First of all, fill in these blanks, writing very plainly:

Name.....Boy or Girl.....

Age.....Grade.....Teacher.....

School.....Date.....

Now do just what the printing tells you to do. There are many questions to be answered. Perhaps you will not be able to answer all of them. That makes no difference. Just do the best you can. Do not go too fast, and try not to make any mistakes. If you come to a question you do not understand, go on to the

6 SCORE	
Arithmetic	
Language	
Reading	
Geography	
History	
Art	
Music	
Morals & Man.	
Phys. Ed.	
Spelling	
Writing	
TOTAL	

[1]After this chapter was written, a volume has appeared which gives a detailed account of a similar series of examinations in Lewis County, New York State. *See* J. S. Orleans and G. A. Sealy, *Objective Tests* (Yonkers-on-Hudson: World Book Company, 1928), pp. x+373. This book, although not a complete treatment of the theory of examination construction, is an exceedingly valuable addition to the practice of objective testing.

next one, and come back to the hard one later on if you have time. There are three kinds of questions. Read each question. If a statement is TRUE, put an(X) in the parentheses. If it is FALSE, put a (−) in the parentheses.

Here is a sample question. Try it for practice. Is the statement true?

Christmas comes in June every other year.....( )

It is false; so the answer should be (−).

Here is another kind of question:

Columbus discovered—..... (.....)

You should write the word America on the line between the parentheses.

Here is another kind of question:

Which floats best on the water? 1. iron; 2. stone; 3. wood; 4. steel; 5. brick. (.....)

The right word is "wood" and a line should be drawn under that word and the number "3" put in the parentheses.

REMEMBER always to put your answer in the parentheses.

Now you are ready to start when the word is given. Do the best you can. You will have one hour in which to finish PART I. Then rest for a while and take another hour for PART II. Do all you can the best you can. You may use scratch paper, if you need it, for any of the Arithmetic work.

PART I

ARITHMETIC. 6

1. $1\frac{9}{12}$ is in lowest terms..... ( )
2. The sum of $\frac{7}{8}$ plus $\frac{1}{2}$ equals $\frac{1}{2}$..... ( )
3. The sum of .08 plus .009 equals .089..... ( )
4. The difference between .20 and .07 equals .13..... ( )
5. $\frac{1}{5}$ equals 25%..... ( )
6. The product of $\frac{2}{3}$ times 8 equals $\frac{1}{12}$..... ( )
7. 25% of 40 equals 10..... ( )
8. 280 is divisible by 3 without remainder..... ( )
9. The difference between $5\frac{1}{4}$ and $2\frac{3}{4}$ equals $2\frac{1}{2}$..... ( )
10. $\frac{1}{4}$ divided by $\frac{3}{8}$ equals $\frac{2}{3}$..... ( )
11. Half of the sum of 6 and 12 equals—(1) 6, (2) 9, (3) 2..... ( )
12. 25¢ is what part of a dollar?—(1) $\frac{1}{4}$, (2) $\frac{1}{2}$, (3) $\frac{1}{3}$..... ( )

13. 7% of \$300 equals (1) $\$42\frac{6}{7}$, (2) \$2100, (3) \$21..... ( )
14. The product of 7.5 times .03 equals—(1) 2.25, (2) .225, (3) 22.5..... ( )
15. 30 divided by $\frac{5}{6}$ equals—(1) 36, (2) 150, (3) 180..... ( )
16. 4.2 divided by 6 equals—(1) 7, (2) .7, (3) .07..... ( )
17. Mr. Sears paid \$100 for a horse, \$12 for a harness and \$48 for a wagon. To find the total cost you should—(1) add, (2) subtract, (3) multiply, (4) divide..... ( )
18. Ruth's cyclometer read 586.7 at starting and exactly 1175 when they reached home. To find how far she rode, you should —(1) add, (2) subtract, (3) multiply, (4) divide..... ( )
19. To find how many yards of ribbon are needed to make a dozen belts each $\frac{3}{4}$ yd. long, you should—(1) add, (2) subtract, (3) multiply, (4) divide..... ( )
20. Lucy reckons that it costs her \$28.75 to feed 15 hens for a year. To find the cost per hen for a year, you should (1) add, (2) subtract, (3) multiply, (4) divide..... ( )
21. Express as a common fraction and write your answer in the parentheses—.75..... ( )
22. Write the missing number in the parentheses—12 equals 50% of. ..... ( )
23. Write the missing number in the parentheses. Traveling for 5 hours at 18 miles per hour you go a distance of miles.....( )
24. What is the cost for one article when you get 10 for \$.25?.....( )
25. Supply the missing number and write your answer in the parentheses—.18 divided by .05 equals ..... ( )
26. A girl buys 3 articles at 75c each and pays \$5. What change should she receive? Write your answer in the parentheses. ( )
27. Write the missing number in the parentheses. A man bought 3 drums for \$1.40, \$1.35, and \$1.00. The average cost was..... ( )
28. John collects rent money for his father. He is paid .02 times what he collects. Write in the parentheses how much he is paid when he collects \$14.50..... ( )
29. Lucy earned \$22.50 in the summer. She spent \$4.25 for books and \$4.05 for clothes. The rest of the money she put in the savings bank. Find the amount she put in the savings bank, and write your answer in the parentheses..... ( )

30. Alice and Helen expect to pick 360 baskets of blackberries this summer and to sell 70 per cent of them. How many baskets do they expect to sell? Write your answer in the parentheses..... ( )

LANGUAGE. 6

1. All sentences should begin with a (1) capital, (2) small letter.. ( )
2. Charles (1) may, (2) can run very fast..... ( )
3. (1) May, (2) can I make a shortcake, if I can find some berries?..... ( )
4. (1) Can, (2) may I have your permission to go?..... ( )
5. Mother (1) may, (2) can I go to visit Mary?..... ( )
6. The apples (1) laid, (2) lay on the ground..... ( )
7. Tom (1) came, (2) come home today..... ( )
8. John (1) sit, (2) set up in your seat..... ( )
9. Please (1) set, (2) sit down..... ( )
10. The abbreviation for January is (1) Jan., (2) Janu..... ( )
11. A peck of peanuts (1) was, (2) were sold..... ( )
12. (1) Has, (2) have each one his book open?..... ( )
13. Call for Clara and (1) me, (2) I..... ( )
14. A bushel of apples (1) were, (2) was sold..... ( )
15. There (1) was, (2) were three of us on the sled..... ( )
16. Each child is in (1) his, (2) their seat..... ( )
17. I shall (1) teach, (2) learn you this trick..... ( )
18. (1) Was, (2) were you there yesterday?..... ( )
19. Will you (1) learn, (2) teach me to play?..... ( )
20. Every one can do this if he (1) tries, (2) try..... ( )
21. The children (1) came, (2) come home tired..... ( )
22. They (1) ran, (2) run all the way..... ( )
23. We did not know they had (1) went, (2) gone until after dinner..... ( )
24. May Ethel and I (1) sit, (2) set together?..... ( )
25. You may (1) sit, (2) set the brown hen..... ( )
26. The brown hen will (1) set, (2) sit on the eggs..... ( )
27. Set the chair over there and (1) set, (2) sit down..... ( )
28. You (1) may, (2) can refer to your books..... ( )

29. Father said I (1) could, (2) might bring Susan to school.... ( )
30. It was (1) they, (2) them who broke the window..... ( )

READING. 6

1. Loki is a character in—(1) Siegfried and the Dragon, (2) The Argonauts, (3) Miles Standish..... ( )
2. The early explorer who said, "Sail on! Sail on! Sail on!" was—(1) Columbus, (2) Drake, (3) the leader of the Pilgrims.... ( )
3. "The Argonauts" was written by (1) Macaulay, (2) Sir Walter Scott, (3) Kingsley..... ( )
4. Jason was the teacher of Chiron..... ( )
5. "The Pied Piper of Hamelin" teaches—(1) It pays to keep a promise, (2) children like music, (3) the river is a good place to drown rats..... ( )
6. "The Moonlight Sonata" was composed by (1) a shoemaker, (2) a blind girl, (3) Beethoven..... ( )
7. "The Inchcape Rock" teaches that—(1) It is fun to sink a bell in the ocean, (2) we suffer for our evil deeds, (3) the Abbot of Aberbrothok was foolish to try to warn sailors by means of a bell..... ( )
8. The "Landing of the Pilgrim Fathers" was written by (1) Hunt, (2) Hemans, (3) Miller..... ( )
9. "Urgan" referred to in "Alice Brand" was a giant..... ( )
10. Herminius is a character in—(1) Horatius, (2) The Cadi's Decision, (3) The Moonlight Sonata..... ( )
11. "They sought a faith's pure shrine," means that they had heard of—(1) a very beautiful altar and were looking for it, (2) they were looking for a place to worship as they believed right, (3) they were coming to America to build cathedrals.. ( )
12. "A Christmas Carol" was written by (1) Browning, (2) Tennyson, (3) Dickens..... ( )
13. Macaulay wrote "Horatius at the Bridge"..... ( )
14. "The Pied Piper of Hamelin" was written by Hemans..... ( )
15. The Pied Piper led the children into the river..... ( )
16. Hamelin was on the river Rhine..... ( )
17. John Maynard forsook the ship and left all to perish..... ( )
18. Southey teaches in "The Battle of Blenheim," that—(1) war always has a glorious victory for some one, (2) the common soldiers always gain most in a war, (3) war is uselessly cruel

and wicked..... ( )

19. In "Horatius at the Bridge" Horatius saved Rome by swimming the Tiber..... ( )
20. A "cadi" is a (1) sheik, (2) judge, (3) beggar..... ( )
21. Jason's men were saved by the sirens..... ( )
22. "The Professor of Signs"—(1) is just a funny story, (2) teaches that a lowly man should keep his temper, (3) shows how easily the same facts may be interpreted in more than one way. ( )
23. Jason sought the Golden Fleece..... ( )
24. King Arthur pulled the sword Excalibur from (1) his scabbard, (2) a stone, (3) a tree..... ( )
25. The poem "Columbus" was written by (1) Coleridge, (2) Miller, (3) Southey..... ( )
26. The story of "A Wonderful City" describes (1) a nest of ants, (2) a city in Asia, (3) a beehive..... ( )
27. The poem "April Rain" was written by (1) Mrs. Browning, (2) Loveman, (3) Longfellow..... ( )
28. "The Inchcape Rock" was written by Wilcox..... ( )
29. Sir Ralph the Rover placed the bell on the rock..... ( )
30. "The Man Worth While" was written by Ella Wheeler Wilcox ( )

END OF PART I. STOP HERE. GO BACK AND SEE THAT ALL OF YOUR ANSWERS ARE RIGHT.

PART II

GEOGRAPHY. 6

1. The wind and rain help to break up the rocks to make soil.... ( )
2. Most of the corn grown in the North Central States is shipped to Europe..... ( )
3. Wheat is the chief crop of the South Central States..... ( )
4. Cotton is raised extensively near Philadelphia..... ( )
5. Pennsylvania is a great coal-mining state..... ( )
6. Rubber is produced from the wood of the bamboo tree..... ( )
7. The region around Para, Brazil, is noted for the production of coffee..... ( )
8. The western part of Brazil is densely populated..... ( )
9. Most of the people of Colombia and Venezuela live on the plateau of the Andes..... ( )

10. Guiana is the only part of South America that belongs to European nations ( )
11. Many cattle are raised on the plateaus of the interior of Brazil ( )
12. Buenos Aires is a city that is larger than New York ( )
13. Quito is the capital of Peru ( )
14. Chocolate is made from a root that is somewhat like a potato ( )
15. South America has a very good system of paved roads and railroads ( )
16. The United Kindgom of Great Britain and Ireland builds more ships than any other country in the world ( )
17. The ruler of the British Empire is chosen in a way similar to the way that the President of the United States is chosen ( )
18. Belgium is a densely populated country ( )
19. Farming is a very important industry in Denmark ( )
20. Coal mining is very important in Norway ( )

HISTORY. 6

1. The most important achievement of early man was learning to live in groups ( )
2. The building of mud and brush huts was also an imporant achievement ( )
3. The use of fire was a disadvantage ( )
4. Flint made better weapons than metal ( )
5. The Egyptians discovered that the year was 365 and $\frac{1}{4}$ days long ( )
6. The Hebrews built large pyramids ( )
7. The Hebrews gave civilization the idea of worshiping just one God ( )
8. The Phœnicians divided the year into months, weeks, days, hours, and minutes ( )
9. The people of the Tigris-Euphrates valley traveled widely and carried with them the learning of Egypt ( )
10. The Greeks gave the world a wonderful literature which has been read and valued by all peoples ( )
11. The greatest help of the Greeks was the idea of the people having a voice in the government ( )
12. The Greeks built wonderful roads through the lands that they conquered ( )

13. Most of our laws are based upon the laws of the Greeks..... ( )
14. Rome made use of the art and learning of Greece..... ( )
15. To the Teutons we owe the idea of each man's being free to express his own belief..... ( )
16. Alfred the Great was a wise king and a good scholar as well.... ( )
17. Alfred the Great was King of—..... (.....)
18. King John was forced to sign the—..... (.....)
19. The were organized to protect the Christian pilgrims who traveled to Jerusalem..... (.....)
20. The was the most powerful man in Europe during the Middle Ages..... (.....)

ART. 6

1. Red, blue, and yellow are primary colors..... ( )
2. Red and blue make—..... (.....)
3. Blue and yellow make—..... (.....)
4. Yellow and red make—..... (.....)
5. Black is a (1) standard color, (2) neutral color..... ( )
6. Red and green make—..... (.....)
7. Orange and blue make—..... (.....)
8. Violet and yellow make—..... (.....)
9. Gray is a neutral color..... ( )
10. White is a neutral color..... ( )
11. Warm colors are restful to the eyes..... ( )
12. Blue, violet, green, gray, and white are cool colors..... ( )
13. Blue, blue-green, and blue-violet are colors..... (.....)
14. Orange, red-orange, and yellow-orange are colors.... (.....)
15. Red, yellow, and orange are warm colors..... ( )

MUSIC. 6

1. A sharp before a note (1) raises, (2) lowers the pitch..... ( )
2. Do and keynote are the same..... ( )
3. Do, mi, sol in any key form what chord?..... (.....)
4. What sign placed before a note lowers its pitch?..... (.....)
5. The sharp farthest to the right in the key signature is always what syllable?..... (.....)
6. The word *ritard* means (1) fast, (2) gradually slower, (3) loud..... ( )

7. The word *forte* means (1) fast, (2) gradually slower, (3) loud..... ( )
8. The word *piano* means—..... (.....)
9. Give the letter which indicates the music is to be played or sung softly..... ( )
10. Give the sign meaning "very loud"..... (.....)
11. The flat farthest to the right in the key signature is always..... (.....)
12. In two measures of 4-4 time there are how many quarter notes?..... (.....)
13. In 6-8 time a quarter rest receives two beats..... ( )
14. There is a tie between two whole notes in 4-4 time; how many measures would you hold the same tone?..... (.....)
15. In 2-4 time an eighth note followed by a dot requires what kind of a note to complete the beat?..... (.....)

MORALS AND MANNERS. 6

1. It often takes great moral courage to tell the truth..... ( )
2. The way boys and girls dress has much to do with making a good impression on other people..... ( )
3. A pupil who throws lunch scraps on the school grounds is helping to make his school attractive..... ( )
4. The comfort of the home depends largely on the helpfulness of its boys and girls..... ( )
5. Girls and boys show kindness by speaking disagreeably about those who are absent..... ( )
6. For boys and girls to develop into good citizens they need (1) merely to avoid breaking the law, (2) merely to learn their lessons at school, (3) to learn all they can at home and at school and to take part in as many as possible of the different forms of group life, such as the playground group, the school club, and the like..... ( )
7. If a young person repeatedly takes little things that do not belong to him, he is (1) really forming the habit of stealing, (2) exercising the right of every person in a free country, (3) doing no harm unless he gets caught..... ( )
8. If a boy is loyal to his gang, (1) he is doing something wrong, (2) he proves he has a splendid quality, (3) he shows a lack of good sportsmanship..... ( )

9. Thrift (1) applies only to putting money in a savings bank, (2) means going without necessary food or clothing, (3) requires careful and thoughtful use of clothes, books, time, money, strength, of all that one has..... ( )
10. If you should find something on the playground that is not yours, you should (1) keep it and say nothing about it, (2) try to find the owner, (3) throw it away..... ( )
11. A pupil is dependable if he behaves well (1) when the teacher is in the room, (2) when a visitor is present, (3) when the pupils are alone in the room..... ( )
12. One way in which to show good school spirit is (1) to be wasteful, (2) to be disrespectful, (3) to be honest..... ( )
13. Why does an employer question a teacher about a boy's or a girl's honesty and truthfulness when he is looking for office help? (1) He is interested in young people, (2) he is interested in the schools, (3) he feels that a boy or a girl who is honest in school will be honest elsewhere..... ( )
14. If another person expresses an opinion different from your own, you should (1) make fun of him, (2) grant him the same right to his opinion that you have to yours, (3) insist that he change his opinion to conform with yours..... ( )
15. If a pupil is poorly dressed, you should (1) tell him you are sorry he is poor, (2) refuse to play with him, (3) pay no attention to his clothes..... ( )

PHYSICAL EDUCATION. 6

In questions 1 to 6 inclusive underline the word that is not related to the other four, and then put the number of that word in the parentheses at the end of the dotted line.

1. (1) baseball, (2) indoor, (3) bat, (4) playground, (5) catcher..... ( )
2. (1) somersault, (2) handspring, (3) stunt, (4) cartwheel, (5) game..... ( )
3. (1) bladder, (2) goal, (3) fish, (4) basketball, (5) jumping ( )
4. (1) sneezing, (2) fresh air, (3) dry feet, (4) sunshine, (5) handkerchief..... ( )
5. (1) health, (2) sleep, (3) vegetables, (4) coffee, (5) toothbrush..... ( )
6. (1) net, (2) reaching, (3) kicking, (4) batting, (5) volleyball..... ( )

7. A ripe banana is a healthful food..... ( )
8. The triple-posture test is a corrective exercise..... ( )
9. The pull-up is a speed-testing event..... ( )
10. The eyes should be guarded from the direct rays of the sun.... ( )
11. Volley-ball is played on a space called a diamond..... ( )
12. Sunshine kills disease germs..... ( )
13. The most important thing about a play-day is (1) the fun of winning, (2) good fellowship, (3) healthful exercise..... ( )
14. Carrousel is (1) a posture test, (2) a running game, (3) rhythmical activity..... ( )
15. Nine-court basketball requires (1) three goals, (2) the playing space divided into nine sections, (3) ten players..... ( )

THE TEST IS OVER. IF YOU HAVE TIME, GO BACK OVER PART II AND MAKE SURE THAT YOUR ANSWERS ARE RIGHT

SPELLING. 6

The teacher will dictate the spelling in sentences.

[Note: Space follows here for writing spelling test from dictation.]

.....

WRITING. 6.

The teacher will score the writing, using Zaner and Bloser, Ayers, or any other good writing scale, allowing 1 to 25 points according to excellence.

EXAMINATION II

The State of Wyoming was one of the first states to employ extensively the objective examination, and is now one of the twenty-odd states administering uniform state-wide examinations. New Jersey has also pioneered in objectifying her state examinations.

The following examination is largely the work of Miss Beatrice McLeod of the State Department of Education, State of Wyoming.[1] It should be noted that this test is designed to cover a range of three grades; in this respect it

[1] *See also* B. McLeod and H. Irving, "Objective Examinations in the Rural Schools of Wyoming," *Journal of Educational Research* Vol. XVIII (1928), pp. 45-49.

resembles the standard test more closely than it does the traditional examination which was almost invariably planned for a single grade.

Page 1—AGRICULTURE

STATE OF WYOMING

State Examination in Agriculture

For Sixth, Seventh, and Eighth Grades

Allow exactly 60 minutes.

Name.....*Grade*.....*Date*.....

School.....*Age*.....*Next Birthday*.....

There are four pages of this test. As soon as you finish one page, go on to the next. Use all the time you have.

TEST I

Directions: If the statement is true, underline the word true.
If the statement is false, underline the word false.
Do not guess. If you are unable to decide whether the statement is true or false, let it alone.

1. A Wyoming farmer does not need an education. TRUE FALSE
2. We should kill as many birds as possible. TRUE FALSE
3. It has been found that sugar beets are a profitable crop in Wyoming. TRUE FALSE
4. A small flock of sheep is of no value to a Wyoming farmer. TRUE FALSE
5. Chicks should be fed immediately after hatching. TRUE FALSE
6. A Jersey cow gives very rich milk. TRUE FALSE
7. All milk should be kept in a sanitary condition. TRUE FALSE
8. It is not necessary to test seed corn. TRUE FALSE
9. Diversified farming is the safest system. TRUE FALSE
10. A good farm organization is of much value to a community. TRUE FALSE

11. All insects are destructive and should be exterminated. TRUE FALSE
12. A hen that is very fat will not make a good layer. TRUE FALSE
13. A dog is no value to a farmer. TRUE FALSE
14. Corn and tomatoes are the principal canning vegetables. TRUE FALSE
15. The people of Wyoming bring their scrub stock to the State Fair. TRUE FALSE
16. The chicken is most important of the fowls. TRUE FALSE
17. Bacteria never work in milk. TRUE FALSE
18. The Merino sheep is noted for his mutton. TRUE FALSE
19. John Burroughs is called "the friend of birds." TRUE FALSE
20. Guano is a valuable commercial fertilizer. TRUE FALSE

TEST II

Directions: In the spaces before the column at the right place the NUMBER of the word that corresponds to the list in the column at the left.

1. Percheron	.....A dairy cow
2. Alfalfa	.....The smallest of living things
3. Ayrshires	.....A baby plant
4. Bacteria	.....External parasite
5. Embryo	.....Carriers of disease
6. Tick	.....Beef type of cattle
7. Rats	.....Leading American crop
8. Hereford	.....A draft horse
9. Corn	.....A breed of hogs
10. Duroc Jersey	.....A leguminous crop

TEST III

Directions: Fill in the blank with the word that makes the best answer.

1. Plants like tomatoes and cabbages should be started in a.....
2. A plant that lives off other plants is called a.....
3. The practice of plowing and cultivating a field one year, in order to grow a crop on it the next year, is called a.....
4. The decayed animal and vegetable matter in the soil is.....

5. Which is the most widely used and most important fiber?.....
6. The principal disease of hogs is.....
7. The danger of this disease has been lessened by.....
8-9. The two main types of hogs are..... .....
10. Feed that supplies ingredients in the proper proportion and amount to meet the needs of the animal is called a.....
11. What is the greatest enemy of the cotton grower?.....
12. Is the earthworm a hindrance or a help to the farmer?.....
13. Who is called the "plant wizard"?.....
14. What large irrigation project is found in Wyoming?.....
15. Where is a large irrigation dam under construction in Wyoming? .....
16. What insect pest is the most numerous in Wyoming?.....
17. What country leads in the production of hogs?.....
18. The most useful insect is the.....
19. Rabbits, prairie dogs, and gophers are very destructive.....
20. What other animal besides the cow produces milk for family use? .....

TEST IV

Directions: Underline one word which completes the sentence.

1. Alfalfa is a kind of corn fruit hay.
2. Bacon comes from the cow hog sheep.
3. The tractor is used in farming mining racing.
4. Rye is most like beans corn wheat.
5. Beets are used for making catsup sugar jellies.
6. Lard comes from butter cattle hogs.
7. A tree that will grow from cuttings is the oak pine willow.
8. The Leghorn is a kind of cow fowl goat.
9. A crop which enriches the soil is clover potatoes tobacco.
10. Milk testers were devised by Babcock Bell Edison.
11. A plant that can be grafted is the apple-tree lily wheat.
12. A good breed of dairy cows is the Holstein Durham Hereford.
13. One of the leading crops in Wyoming is rice tobacco hay.
14. The soil is enriched by osmosis propagation legumes.
15. We get rid of the potato bug by spraying fumigating dipping.

TEST V

Directions: Put a cross (X) before the best answer.

1. We cull our flock of hens in order that we may have:
   A better looking flock.
   A better laying flock.
   A smaller flock.

2. Crop rotation is practiced by farmers in order to:
   Lengthen the period of fertility of the land.
   Adapt the crops to the season.
   Prevent the land from lying idle through the winter.

3. Leguminous plants are important because:
   They grow on soil too poor to support other crops.
   They return nitrogen to the soil.
   They are easily cultivated.

4. A Wyoming farmer cultivates his corn soon after a rain:
   To enrich the soil.
   To compress the soil firmly about the roots.
   To form a mulch and check evaporation.

5. A farmer should protect the birds:
   Because of their beautiful plumage.
   Because of their sweet music.
   Because they destroy many harmful insects.

TEST VI

Directions: Write on each line the word or words which complete the sentences. Don't waste too much time on one you do not know. Go on and come back to it later.

1. The parts of a plant are (1)....., (2)....., (3)....., (4)....., (5).....
2. We cultivate a field to (6).....
   (7).....
   (8).....
   (9).....
   (10).....

3. The different kinds of soil are (11)....., (12)....., (13).....
4. Some of the leguminous crops in Wyoming are (14)....., (15)....., (16).....
5. Three important dry-farming crops are (17)....., (18)....., (19).....
6. The four good "general purposes" breeds of chicken are the (20) ....., (21)....., (22)....., (23).....
7. The two best-known egg-producing breeds are (24)....., (25).....
   The largest meat-producing breed is the (26).....
8. Besides chickens the Wyoming farmer raises many large flocks of (27)....., (28).....
9. Weeds are harmful because they not only (29)....., but often (30).....

EXAMINATION III

Examination III is the work of Mr. Sam Everett and Miss Effey Riley of East High School, Rochester, New York. This test was constructed in January, 1928, and was revised the following June, the revision being reprinted here by permission of the Rochester Board of Education. Note that this examination combines old and new types of testing. (See Exercise IV for "essay" question and Exercise VI for use of controlled "short-answer" or association technique.)

EAST HIGH SCHOOL

AMERICAN HISTORY

Term

(HISTORY III-1)

June, 1928

GENERAL DIRECTIONS:

This examination has been carefully planned in order to test the different kinds of abilities in which we have tried to give you some training. Do not hurry. It is far more important to try and answer each exercise carefully than to try to finish. Answer the questions in order. Directions are given with each exercise.

EXERCISE I. To test both your knowledge of the relationship of a number of great Americans to certain historical periods and of their significance within their own period.

A. From the list at the right of the page, select the names of five people in each of the following periods. Write their names under the name of the period in which they were prominent. Note that some names may be used more than once and that there are some names which may not be used at all.

1. The Colonial Period (1607-1763)

.....
.....
.....
.....
.....

2. Men who helped to form the U. S. Constitution and favored its acceptance by the States.

.....
.....
.....
.....
.....

3. Leading men in sympathy with the Federalist Party and with Federalist ideals.

.....
.....
.....
.....
.....

4. Leading men in the early Republican Party.

.....
.....
.....
.....
.....

George Washington
Roger Williams
John Cabot
John Calhoun
John Adams
Governor Clinton
Patrick Henry
Ann Hutchinson
Alexander Hamilton
Robert Morris
Peter Stuyvesant
Henry Hudson
John Marshall
John Jay
Henry Clay
James Madison
William Penn
James Monroe
Thomas Jefferson
John Winthrop
Andrew Jackson

B. Indicate after each name two significant historical facts which come to your mind in connection with the man in question.

1. Robert Morris.....

2. William Penn.....

3. Alexander Hamilton.....

4. John Marshall.....

5. Thomas Jefferson.....

6. Roger Williams.....

EXERCISE II. To test your knowledge of important geographical facts of colonial times that have affected the development of American civilization.

Place the number of the best answer on the line provided at the right of each statement.

A. The great natural gateway from the Atlantic coast into the West that Americans knew must be held if the revolution of the 13 colonies was to be permanently successful was ..... (1) Chesapeake Bay, (2) Connecticut Valley, (3) Hudson Valley, (4) James River, (5) St. Lawrence River.

B. The rise of manufacturing in New England was greatly aided by the fact that their physical environment furnished ..... (1) cold temperature, (2) all kinds of raw materials, (3) many navigable rivers, (4) easy communication with the West, (5) water power.

C. The bulk of trade of the Colonial South was with ..... (1) New England, (2) England, (3) West Indies, (4) Spain, (5) Different Southern Colonies.

D. The most important trade of the New England colonies, before the American Revolution, was with ..... (1) South America, (2) West Indies, (3) Spain, (4) Far West, (5) England.

E. The principal crop of Virginia in colonial times was .....
(1) wheat, (2) tobacco, (3) potatoes, (4) cotton, (5) furs.

F. The industry that brought to Colonial Massachusetts the greatest prosperity was .....
(1) potatoes, (2) corn, (3) hemp, (4) fish, (5) hats.

G. The wealth of Colonial South Carolina came chiefly from .....
(1) rice, (2) tobacco, (3) cotton, (4) furs, (5) wheat.

H. The character of soil, climate, and location allows the development of a great variety of agricultural products in .....
(1) New England, (2) Middle States, (3) Virginia, (4) Appalachian Mt. Region, (5) Georgia.

EXERCISE III. To test your ability to recognize clearly conditions of social environment of different periods of American history.

Each of the statements below can be completed by one of the five different numbered phrases. Read each statement. Decide which of the numbered phrases, when added to the original statement, will make it true and complete. Then place the number of the completing phrase on the dotted line at the right of the statement.

A. During the colonial period democracy was .....
(1) common in all the colonies, (2) confined to Rhode Island, (3) everywhere limited by property qualifications, (4) stamped out by the Crown, (5) found only in Massachusetts as a result of the Mayflower Compact.

B. In colonial times voting was the privilege of .....
(1) all the inhabitants, (2) all male whites over 21 years, (3) those born or naturalized in the U. S., (4) all persons who could read or write, (5) those who possessed either religious or property qualifications, or both.

C. In the Middle Colonies the dominant character developed was .....
(1) Yankee, (2) Puritan, (3) Planter, (4) Poor white, (5) Quaker.

D. The character of the Puritan may best be described as .....
(1) delightful, easy-going, aristocratic, (2) friendly, equality-loving, thrifty, peaceful, (3) strict, thrifty, deeply religious, intolerant, conscientious, (4) rough, self-reliant, adventuresome, (5) shiftless, poor, ignorant.

E. The Middle Colonies were peopled mainly by .....
(1) English, Scotch-Irish, and Welsh, (2) French, Germans, English, (3) French, English, (4) English only, (5) German English, Swedes, Dutch.

F. The Appalachian chain of mountains was chiefly significant in our early colonial history because .....
(1) it sheltered Indian marauders, (2) it hindered the westward advance of the colonists, (3) its fine timber was used for ship building, (4) it contained rich iron and coal deposits, (5) it contained rich grazing lands.

G. "The good education of children is of singular benefit to any community" was first declared an American ideal by .....
(1) Virginia, (2) New York, (3) Pennsylvania, (4) Massachusetts, (5) Georgia.

H. The American ideal of equality was a most nearly realized fact .....
(1) in New England, (2) with the Dutch in N. Y., (3) on Southern plantations, (4) on Western frontier, (5) in the Quaker settlements of Eastern colonies.

I. Religious toleration in colonial times was found in .....
(1) Virginia, (2) Massachusetts, (3) New Hampshire, (4) Maryland, (5) Georgia.

J. At the time of the adoption of the Constitution the states were .....
(1) all in favor of the new constitution, (2) anxious for separate constitutions, (3) in favor of it except the New England colonies, (4) within two years mostly in favor of it, (5) never in final agreement concerning it.

K. By the time of Jefferson's first inaugural, voting privileges in most of the States were .....
(1) extended to all, (2) restricted to manhood suffrage, (3) still limited by property and religious restrictions, (4) limited only by religious restrictions, (5) given to many Indians who had become naturalized.

L. The chief significance of the capture of the government by the Jacksonian Democrats as we see it today was that .....
(1) it brought an end to "The Era of Good Feeling," (2) it gave western politicians control of its government, (3) the National Bank was abolished, (4) Jackson was a great Indian fighter, (5) many democratic institutions such as white manhood suffrage came to be established.

EXERCISE IV. To test your ability to see how an intelligent knowledge of past events helps us to understand present-day situations, and tendencies.

(Note: Write your answer in essay form on a separate sheet of paper.)

Some one has said that we study the past relationships in American life in order to be able to understand the present in our civilization and that we need to understand the present so as to influence American national development toward finer things.

State your reasons for every position assumed.

a. Take some *economic* fact or group of facts in American History about which we have studied and briefly show what seems to you to be the actual significance of this fact in the past, present and future of America.
b. Show this same *three-fold relationship* using some *political* fact or facts.
c. Show this same three-fold relationship using a *religious* fact or facts.

EXERCISE V. To test your ability to recognize some of the precious social heritages that have come down to our present-day America from the past.

A. What do we mean by "social heritage"?

[EDITOR'S NOTE: In the original, sufficient space was allowed here for writing the response to question A.]

B. Below is a list of statements. Indicate by a cross (X) after it, each statement that expresses a social heritage of the present-day American nation.

Place a (0) after each statement that is not a present-day social heritage of the American nation.

1. Americans believe in the ideal of religious toleration. .....
2. Property in land should be inherited by a man's eldest son. .....
3. Citizens should have the right to say what taxes should be put upon them. .....
4. No man's house shall be searched for evidence of law violation unless the searchers have a written permit stating exactly what they are searching for, and who made the charge of law violation. .....
5. The majority of citizens shall always have the right to state what shall be the religious faith practiced in the community. .....
6. Government should interfere with men's lives and freedom as little as possible. .....
7. An ideal of society is a belief in the union of church and state. .....

8. States may have many sovereign powers, and still be obedient to some higher central authority in some common matters that affect several states alike. .....
9. Government shall be separated into three powers: executive, legislative, judicial. .....
10. An aristocratic class is necessary in order that a nation shall have fine, intelligent, moral and cultural leadership. .....

EXERCISE VI. To test your ability to recognize and judge the significance of certain famous events in our national history.

Indicate by a descriptive sentence your *exact* historical knowledge of each of the following, and its significance.

[EDITOR'S NOTE: In the original, space for the response was allowed after each item.]

1. The introduction of tobacco culture
2. The founding of Pennsylvania
3. The invention of the cotton gin
4. The Philadelphia Convention
5. The American Bill of Rights
6. The Lewis and Clark Expedition
7. Jefferson's first inaugural
8. Jackson's first administration

EXERCISE VII. To test your ability to recognize political theories and opinions of prominent party leaders.

The following is a list of theories held either by Alexander Hamilton or Thomas Jefferson. Each theory is numbered. Place the number of each theory either after the name of Jefferson or after the name of Hamilton depending on whether you feel it was believed by the one or the other. Note that certain theories may be held by both men.

Hamilton..... Jefferson.....

1. There should be a strong central government in the United States.
2. The national government should pass laws which should chiefly benefit the poor.
3. There was no need in the Constitution of guaranteeing certain "inalienable rights" to the people.
4. Poor people are likable and can be trusted.
5. The national government should spend money on internal improvements, especially on roads into the west.

6. The United States should become an industrial nation.
7. The states should not give up many of their rights in order to strengthen the national government.
8. The national government's credit and position would be best strengthened by decreasing the national debt.
9. The United States should become an agricultural nation.
10. "Your people is a great beast."
11. The French Revolution should be distrusted and condemned.
12. The national government should pass laws which should principally benefit the rich.
13. The frontier sections of the country are distrusted, and there is refusal to aid these people with national government funds.
14. It would strengthen the national government to assume certain former national and state debts.
15. It would be best for the United States government in the long run not to pay any bribes in order to protect our commerce.
16. The French Revolution was a splendid movement and made for the betterment of mankind.
17. The Bill of Rights should be strictly enforced.

EXERCISE VIII. To test your ability to see and explain different ways in which environment and people may affect each other.

(Note: Write your answer on a separate sheet of paper.)

Most people think that the immediate environment in which we live largely determines the ideas of the majority of us, and that that environment, so far as each of us is concerned, is apt to be quite accidental.

a. Do you think this statement is true as it applies to ordinary people in such distinct periods of our country's history as Puritan New England, Colonial Pennsylvania, or our own generation? Give illustrations or factual evidence on which you base your opinion.

b. Is it possible for individuals to have any moral or intellectual standards independent of their immediate environment? Discuss this, basing your opinion on facts taken from the lives of such men as Roger Williams, Benjamin Franklin, Thomas Jefferson, or any other men that you care to take.

c. Is the attempt at control or change of one's environment important or not for the life of each of us? On what factual evidence do you base this judgment?

d. If it could be attained, what do you think would make the best kind of social environment for the lasting success and happiness of each of us? Would this apply equally to America as a whole?

EXERCISE IX. To test the opinions you have formed on certain facts in American history.

Below are listed various statements about early American history. Draw a circle around the letter or question mark which best indicates the way you feel about each statement, as follows:

(R) ? W If you have a feeling in favor of the statement, draw a circle around R.

R ? (W) If you have a feeling against the statement, draw a circle around W.

R (?) W If you are quite uncertain as to knowledge or feeling, draw a circle around the ?.

Mark every item. Omit none. If you do not understand any item, simply put an X before the item.

R ? W 1. The Puritans and Pilgrims came to America because of their suffering from religious persecution at home. They, therefore, determined to make religious toleration a cornerstone of their religious beliefs and of their government in the new world.

R ? W 2. American colonists in general early welcomed and treated as equals all oppressed peoples, including Jews and Catholics.

R ? W 3. None of the colonies in colonial times believed in and practiced religious toleration.

R ? W 4. The German settlers coming to our country in colonial times made the poorest settlers because they were both ignorant and militaristic.

R ? W 5. White men, as well as black men, were enslaved in early America.

R ? W 6. Slavery never flourished in the Northern colonies because the institution of slavery was against the Christian ideals of the Puritans.

R ? W 7. Colonial life in America after the first twenty-five years of settlement was pleasant and easy for the majority of the colonists.

R ? W 8. The Quakers were very unpopular with the Puritans of the New England colonies because of their religious beliefs.

R ? W 9. Colonial trade and industry became so large that it excited the fears and jealousies of English competitors.

R ? W 10. The Pilgrim Fathers were kind to everyone no matter what their religious beliefs.

R ? W 11. Among the free inhabitants of America until after the Revolution, social life was upon a basis of almost absolute equality.

R ? W 12. Almost without exception, the people coming to America in colonial times represented the very best stock (physical, mental, social) of the countries from which they came.

R ? W 13. Before the Revolution the rough life of the colonies had prevented them from making any notable contributions to science or other branches of learning.

R ? W 14. At least one-half of the immigrants in America before the Revolution were slaves or bond servants.

R ? W 15. The Indians with whom the colonies carried on warfare were extremely ferocious, and practically always their attacks on the settlements were sudden and unjustified.

R ? W 16. The Quakers of Pennsylvania had no colonial militia with which to overawe the Indians; and yet, an Indian uprising against them was comparatively unknown.

R ? W 17. Rhode Island, founded by Roger Williams on the ideal of religious and political freedom, was from the first one of our most orderly and successful colonies.

R ? W 18. The majority of the settlers coming to America were Puritans, and for this reason their ideals became American ideals.

R ? W 19. Since a great number of the colonists had come to America for political freedom and to found governments on democratic ideals, full manhood suffrage was granted in every colony from the first.

R ? W 20. The major reason why slavery did not flourish in the New England colonies was because it was not a good financial proposition.

R ? W 21. Before the Revolution there had grown up dissension and strong feelings of antagonism between certain sections of the frontier and the Atlantic coast communities.

R ? W 22. The democratic town meeting was the typical type (almost universal) of local government in the thirteen colonies.

R ? W 23. The American colonies before the Revolution presented a "strange mingling of the uncouth, the totally wild, and the highly civilized and cultured."

R ? W 24. An efficient and powerful national government was set up under the Articles of Confederation.

R ? W 25. The National Government under the Articles of Confederation could do nothing to suppress popular disorders and rebellions.

R ? W 26. The men who wrote the Constitution drew it up in secret session where the public could know nothing of what was going on.

R ? W 27. The delegates to the Constitutional Convention at Philadelphia with a few unimportant exceptions tried their best to establish the new government on the broadest democratic basis possible.

R ? W 28. The delegates to the Constitutional Convention at Philadelphia were more concerned in having clauses which protected private property than those protecting individual freedom.

R ? W 29. Many of the members at this Convention had at one time been leaders in rebellion against law and authority.

R ? W 30. Alexander Hamilton as a member of the Convention wholeheartedly submitted a plan of government that proposed a monarchy, with Washington as first king.

R ? W 31. The Constitution represents a series of compromises rather than a document considered perfect by its signers.

R ? W 32. Alexander Hamilton said of the Constitution that it was a "flimsy" document, that "would not last a year."

R ? W 33. Some members of our Constitutional Convention were influenced by their private or class interests in drawing up certain parts of the Constitution.

R ? W 34. While the Constitution was being drawn up sectional jealousies frequently divided the delegates.

R ? W 35. After the drawing up of the Constitution it was ratified by the various states with little or no opposition.

R ? W 36. The supremacy of the U. S. Constitution as the ultimate authority over all the people has never been seriously questioned since its adoption.

R ? W 37. Intelligent people who study the Constitution in detail, as it works out in practice, feel that the founders gave to our people a nearly perfect document that will never need much change.

R ? W 38. The Constitutional Convention was peaceful and encountered no difference in issues.

R ? W 39. Our Constitution represents a new idea in government, created by the thought of the brilliant men of genius who formed our Constitutional Convention.

R ? W 40. Freedom of speech, of the press, of assemblage, of religion were guaranteed in the Constitution at the time it was originally drawn and first presented to the people for ratification.

EXERCISE X. To test whether or not the study of American History has meant to you mere "book-learning."

(Note: Write your answer on a separate sheet of paper.)

A. Carefully explain the meaning of "book-learning" as it is used in the following quotation:

"Education is not *book-learning.* It has to do with insight, with valuing, with understanding, and with the development of the ability to make a choice among the possibilities of experience."

B. As part of your education you have been studying in American history about the Constitutional Convention. Has the study of that historical event meant to you simply memorizing a list of facts or events,—or has it given you (1) insight into the significance of certain decisions made by the men of the Constitutional Convention; (2) ability to evaluate certain clauses of our Constitution; (3) ability to decide whether our forefathers intended to give us a democracy, or not?

If you have gained any of these three things, will you try to show that you have acquired them through use of practical illustrations in each of the three cases?

EXERCISE XI. To test your ability to reason clearly using historical facts and truths as a basis.

(Note: Write your answer on a separate sheet of paper.)

A. From your study of American History illustrate the probable truth of the following statements by comparing several earlier periods with our own.

"In this very uncertain world of ours, ways of living, standards of value, customs, and traditions are always in process of continual change."

B. Recognizing this truth, what do you think that an education should do for each one of us? Should we be taught what to think, or how to think? Use a number of illustrations in thinking through this problem. Make sure, in your answer, that you show that you understand the difference in meaning between the two phrases, "what to think," and "how to think."

EXAMINATION IV

Examination IV is one of a series of objective tests produced in connection with the Summer Library Institute of the American Library Association held at the University of Chicago during the summer of 1926. The complete series includes tests on the following library-school subjects: Book Selection, Reference Work, Library Classification, Lending Methods, School-Library Administration, Children's Work, and How to Use the Library. The last mentioned is given here, with omissions, as being the one of greatest interest to teachers other than librarians. It is the work of Miss Linda M. Clatworthy, Miss Sadie T. Kent, Miss Anna C. Lagergren, and Miss Delia V. Ovitz. General supervision of the construction of these tests was given by the instructors in the Summer Library School, particularly Professor Sidney B. Mitchell and the present author.

Although these tests are not generally available, information concerning a limited mimeographed edition of the same may be had by addressing the American Library Association, 86 East Randolph Street, Chicago, Illinois.

OBJECTIVE EXAMINATION

Undergraduate Course: "How to use the Library"

TRUE-FALSE

1. The *Standard Dictionary*, in defining a word, gives the literal or original meaning first. TRUE FALSE
2. The *New International Dictionary* gives the common meaning first. TRUE FALSE
3. The definitions in the *New International Dictionary* are fuller than the definitions in the *Century Dictionary*. TRUE FALSE
4. The system for showing the pronunciation is the same in all the dictionaries. TRUE FALSE
5. The etymologies are fullest in the *Century* and briefest in the new *Standard*. TRUE FALSE
6. For abbreviations and foreign words and phrases, the *New International* or the *New Standard* is better than the *Century*. TRUE FALSE

7. The *New International Dictionary* gives antonyms. TRUE FALSE
8. All three dictionaries give synonyms. TRUE FALSE
9. For proper names, the fullest treatment is given in the *Century*. TRUE FALSE
10. The pages of the *New Standard* are divided into two sections; the words not in common use are put in finer print in the lower section of the page. TRUE FALSE
11. *Webster's Dictionary* gives the fullest account of the history of a word. TRUE FALSE
12. *Oxford Dictionary* is another name for *Murray's New English Dictionary*. TRUE FALSE
13. The two supplementary volumes of the *Century Dictionary* published in 1909 were incorporated in the 1911 edition in alphabetic order. TRUE FALSE
14. The information on any subject is so scattered in the *New International Encyclopedia* that in order to be sure you have it all, you must consult the index. TRUE FALSE
15. The *Britannica* has very full cross-references. TRUE FALSE
16. The *Americana* is stronger in scientific and technical material than the *New International*. TRUE FALSE
17. You will find the best treatment of the American Revolution from the American viewpoint in the *Britannica*. TRUE FALSE
18. The material in *Nelson's Encyclopedia* is kept up to date by a Yearbook. TRUE FALSE
19. The alphabetical arrangement in the *New International* is letter by letter instead of word by word; for instance, New Jersey, Newspaper, New York. TRUE FALSE
20. The *Americana Encyclopedia* and the *New International* cover much the same ground. TRUE FALSE
21. The *Americana* is the only loose-leaf encyclopedia. TRUE FALSE
22. All encyclopedias follow the same scheme of arrangement of their subject-material. TRUE FALSE
23. The articles printed in the *New International Encyclopedia* are all signed. TRUE FALSE
24. The *Encyclopedia Britannica* is the most concise and comprehensive of all the encyclopedias. TRUE FALSE
25. The index to the main part of the *New International Encyclopedia* is in Vol. 24. TRUE FALSE
26. The Harvard Classics are published in ten volumes. TRUE FALSE
27. The Harvard Classics are often spoken of as Eliot's Five-foot Book Shelf. TRUE FALSE

28. *Whitaker's Almanac* contains tables and lists, chiefly applicable to Great Britain; statistics and information about the government of all countries. TRUE FALSE
29. You can find the duties of a department of the U. S. government in the *U. S. Congressional Directory.* TRUE FALSE
30. A biography of Abraham Lincoln will be found in *Who's Who in America.* TRUE FALSE
31. The material in the *Bartlett's Familiar Quotations* is arranged alphabetically by subject. TRUE FALSE
32. In Hoyt's *Cyclopedia of Poetical Quotations* all quotations from a given author are in one place. TRUE FALSE
33. The *Publisher's Weekly* gives you the author, title, publisher, and price of books published during the week. TRUE FALSE
34. If you wish a list of the important books of the year published in America on any subject consult the *Cumulative Book Index* for that year. TRUE FALSE
35. The *Book Review Digest* has a subject and title index. TRUE FALSE
36. The *World Almanac* is published bi-annually. TRUE FALSE
37. The Warner Library consists of extracts from the literature of all countries. TRUE FALSE
38. The *Book Review Digest* is kept up to date by a monthly supplement. TRUE FALSE
39. The *Statesman's Yearbook* is devoted entirely to descriptions and statistics of the governments, industries, and resources of the U. S. TRUE FALSE
40. The *Statesman's Yearbook* is made up of long signed articles by specialists. TRUE FALSE
41. *Poole's Index* is a guide to magazine articles since 1900. TRUE FALSE
42. The *Reader's Guide* covers the 19th century. TRUE FALSE
43. The *Reader's Guide* is published monthly. TRUE FALSE
44. The *Reader's Guide* indexes all of the important magazines published in the U. S. and foreign countries. TRUE FALSE
45. Inclusive pages are given for the articles in *Poole's Index.* TRUE FALSE
46. If the library does not have a bound volume of Robert Frost's poems, such poems as have appeared in magazines the past few years may be located through *Poole's Index.* TRUE FALSE
47. A contemporary review of Scott's *Lady of the Lake* may be located through the *Reader's Guide.* TRUE FALSE
48. Good debate material on the commission form of government may be secured from the *Reader's Guide.* TRUE FALSE

49. The "Library of Congress" scheme is the scheme of classification most frequently used by libraries. TRUE FALSE
50. The Dewey scheme of classification was devised by John Dewey, the psychologist. TRUE FALSE

. . . . .

108. Full bibliographic information about books and articles referred to in the text can usually be found in footnotes or bibliography at end of chapter or book. TRUE FALSE
109. A rather full outline of a book may sometimes be found in the table of contents. TRUE FALSE
110. Reading the preface of a book sometimes helps get the purpose for which the book can be used. TRUE FALSE
111. "Q" indicates book is larger than "F." TRUE FALSE
112. In a dictionary catalog the subjects are usually in red. TRUE FALSE

MULTIPLE-RESPONSE

113. The Articles of Confederation can be found in
    1. New International Yearbook
    2. Statesman's Yearbook
    3. Harper's Encyclopedia of U. S. History
    4. Hart and McLaughlin—Cyclopedia of American Government
    5. World's Almanac

114. Sketches of living Americans can be found in
    1. National Dictionary of Biography
    2. Lippincott's Biographical Dictionary
    3. Who's Who in America
    4. Appleton's Cyclopedia of American Biography
    5. Century Cyclopedia of Names

115. The origin of famous names in fiction may be found in
    1. Baker's Guide to Best Fiction
    2. Brewer's Readers Handbook
    3. Statesman's Yearbook
    4. Readers' Guide
    5. World Almanac

116. The author and source of poetry and recitations may be found in
    1. Firkins—Index to Short Stories
    2. Granger—Index to Poetry and Recitations
    3. Book Review Digest
    4. Ward—English Poets
    5. A. L. A.—Index to General Literature

117. The leading articles in all the important periodicals are indexed in the
    1. Cumulative Book Index
    2. Book Review Digest
    3. Reader's Guide to Periodical Literature
    4. Card Catalog
    5. Gurrance—Guide to Periodicals

118. The Statistical Abstract of the U. S. contains
    1. Statistical tables of the last census
    2. Abstract of deeds
    3. Current statistics
    4. Statistics about government
    5. Statistical maps of the last census

119. In which book would you look to find a review of Willa Cather's *The Lost Lady?*
    1. Poole's Index
    2. Book Review Digest
    3. Cumulative Index
    4. A. L. A. Book List
    5. Reader's Guide

. . . . .

157. If you wish to find a poem and can remember the first line, what reference book would you consult?
    1. Hoyt's Cyclopedia of Quotations
    2. Granger's Index to Poetry and Recitation
    3. Carman—World's Best Poetry
    4. Dana—Household Book of Poetry
    5. Quiller-Couch—Oxford Book of English Verse

158. For pronunciation of places, where would you look?
    1. Rand McNally—Commercial Atlas
    2. Lippincott's Gazeteer
    3. Century Atlas
    4. World Almanac
    5. New International Encyclopedia

159. For concise articles on English history consult
    1. Ploetz—Epitome of Universal History
    2. Low and Pulling—Dictionary of English History
    3. Larned—History for Ready Reference
    4. Brewer—Historic Notebook
    5. Heilprin—Historical Reference Book

160. Where would you find briefs and reports of intercollegiate debates on present-day questions?
   1. World Almanac
   2. Congressional Directory
   3. University Debaters' Annual
   4. Wilson Debaters' Handbook Series
   5. Matson—Reference for Literary Workers

161. Popularly written, yet scientifically authentic articles on any phase of agriculture can be found in
   1. Bailey—Cyclopedia of Agriculture
   2. Yearbook of Agriculture
   3. Statesman's Yearbook
   4. American Yearbook
   5. World Almanac

162. Where may you find authentic information on topics of current educational interest?
   1. New International Yearbook
   2. Statesman's Yearbook
   3. World Almanac
   4. Book Review Digest
   5. U. S. Bureau of Education Bulletins

Matching Exercises

Classification: Dewey Decimal

163.

	No.	*Classification*	*Ans.*
1.	100	General Works	(.....)
2.	200	Sociology	(.....)
3.	300	Religion	(.....)
4.	400	Philosophy	(.....)
5.	500	Natural Science	(.....)
6.	600	Fine Arts	(.....)
7.	700	Philology	(.....)
8.	800	Useful Arts	(.....)
9.	900	Literature	(.....)
10.	000	History	(.....)

164.

1.	370	Botany	(.....)
2.	822	French grammar	(.....)
3.	780	American history	(.....)

4.	811	Economics	(.....)
5.	150	Zoölogy	(.....)
6.	840	Psychology	(.....)
7.	580	Geology	(.....)
8.	973	Agriculture	(.....)
9.	330	Home economics	(.....)
10.	640	Printing	(.....)
11.	750	American poetry	(.....)
12.	590	Music	(.....)
13.	445	English drama	(.....)
14.	630	French literature	(.....)
15.	550	Education	(.....)

Recall Questions

165. Consult the catalog for author entry of the *Proceedings of the American Library Association* under .....
166. Consult catalog for the author entry of the *Report of the Mass. Dept. of Education* under .....
167. Consult catalog for author entry of our Federal Bureau of American Ethnology under .....
168. Consult catalog for the *Confessions of St. Augustine* under .....
169. The daily record of the debates and business of Congress is called the .....
170. The chief book form of ancient Babylonia was the .....
171. The chief book form of ancient Egypt was the .....
172. The early book form of ancient Greece and Rome was the .....
173. "Il." in the catalog means .....
174. "Por." in the catalog and periodical indexes means .....
175. "Pl." in the catalog means .....
176. "Enl. ed." in the catalog means .....
177. "Rev. ed." in the catalog means .....
178. "5 v." in the catalog means .....
179. Write out this abbreviation: 67th Cong. 4th. sess. H. Doc. 323 .....
180. The leading index to newspapers is the .....
181. The symbols found in the *Reader's Guide*, such as Lit. Dig. 47:25-6 Je 15'24, mean .....
182. *Ibid* means .....
183. "cl926" means .....

184. "q.v." means .....
185. "do." means .....
186. The modern successor to *Poole's Index* is .....
187. In the field of foreign periodicals consult the .....
188. In the field of technology periodicals consult .....
189. In the field of agriculture and home economics periodicals consult .....
190. In the field of business periodicals consult .....
191. The *Reader's Guide* began indexing periodicals in .....
192. Good check-lists of periodicals may be found in front of .....
193. In the *Reader's Guide* a single poem (author not recalled) may be found under .....
194. In the *Reader's Guide* short stories (author not recalled) may be located under .....
195. Book-review sections of magazines are indexed in the magazine indexes up to .....
196. After above date, book reviews may be found in .....

EXAMINATION V

Examination V shows a most interesting and original approach to a very difficult type of measurement, viz., the appreciation of the qualities of literature. This test was constructed by Mr. Arthur Agard of the Alameda, California, High School. This examination won first prize in a competition with nearly one hundred objective tests in English. It also tied for first place among 375 entries representing eight principal groups of high-school subjects.[1]

Mr. Agard's work represents a continuation of the line of development in the testing of literary qualities begun by Dr. M. F. Carpenter of the State University of Iowa.[2]

In the opinion of the author, the work of Agard and Carpenter represents one of the significant contributions to the technique of the objective measurement of the teaching of literature in the high school and college.

[1] *See* G. M. Ruch and G. A. Rice, *Specimen Objective Examinations.*
[2] *Improvement of the Written Examination,* pp. 86-90.

OBJECTIVE TEST
ON THE QUALITIES OF A PASSAGE IN LITERATURE

TEST I: IDENTIFICATION TEST FOR ONE QUALITY—20 POINTS

Each of the following passages is especially noteworthy for some *one* of the following qualities:

A. Skillful phrasing (compactness, the exactly right word, wording could not be changed without weakening)
B. Adaptation of sound to meaning (mimetic words; appropriate rhythms; use of mutes, gutterals, aspirates, liquids, long vowels to produce desired effects)
C. Beauty of image (attractive because of desirable emotion, memory, or imagination appeal)
D. Force of image (definiteness, unusualness, vividness, striking to imagination, many points of likeness in figures)
E. Worth of thought

Place the letter for the quality at the left of the first line of the passage.

The student is advised, in case he recognizes the source of the quotation, to consider the selection here given only.

The student is advised to test each passage for each quality, and by the Method of Residues[1] to eliminate all possibilities save the one quality finally determined.

..... 1. As rivers of waters in a dry place,
As the shadow of a great rock in a weary land.

..... 2. As for the grass, it grew as scant as hair in leprosy.

..... 3. And now the sun has stretched out all the hills.

..... 4. And ten low words oft creep in one dull line.

..... 5. His honor rooted in dishonor stood,
And faith, unfaithful, kept him falsely true.

..... 6. Battle's magnificently stern array.

..... 7. Let us sit, while my mind remembers
The beauty of fire in the beauty of embers.

..... 8. The old order changeth, yielding place to new
And God fulfills himself in many ways.

..... 9. Million-footed Manhattan, unpent, descends to her pavements.

.....10. Or else, as if the world were wholly fair,
But that these eyes of men are dense and dim
And have not power to see it as it is—
Perchance because we see not to the close.

[1]The "Method of Residues" refers to successive eliminations.

.....11. Her eyes were deeper than the depth
Of water stilled at even—

.....12. The cataracts blow their trumpets from the steep.

.....13. It's coming yet for a'that
That man to man, the world o'er,
Shall brothers be, for a'that.

.....14. A chuckle of laughter like the tapping of unstrung kettledrums

.....15. Magic casements opening on the foam
Of perilous seas, in faery lands forlorn.

.....16. The knight's bones are dust,
And his good sword rust;
His soul is with the saints, I trust.

.....17. Little flower,—but if I could understand
What you are, root and all, and all in all
I should know what God and man is.

.....18. Without a word of warning, there
In the autumn sky Mount Fuji stands.

.....19. A solitary shriek, a bubbling cry
Of some strong swimmer in his agony.

.....20. He rushed into the field, and foremost fighting fell.

Test II: Identification Test for Two Qualities—20 Points

Each of the following passages is especially noteworthy for *two* of the following qualities:

A. Skillful phrasing (compactness, the exactly right word, wording could not be changed without weakening)
B. Adaptation of sound meaning (mimetic words; appropriate rhythms; use of mute, gutterals, aspirates, liquids, long vowels to produce desired effects)
C. Beauty of image (attractive because of desirable emotion, memory, or imagination appeal)
D. Force of image (definiteness, unusualness, vividness striking to imagination, many points of likeness in figures)
E. Worth of thought
F. Effective contrast of main ideas

Place the letters for the two qualities at the left of the first line of the passage.

The student is advised, in case he recognizes the source of the quotation, to consider the selection here given only.

The student is advised to test each passage, for each quality, and by the Method of Residues to eliminate all possibilities save the two qualities finally determined.

..... 1. The Rank is but the guinea's stamp,
The man's the gold for a' that.

..... 2. Though I speak with the tongues of men and of angels and have not charity, I am become as sounding brass or a tinkling cymbal.

..... 3. The league-long roller thundering on the reef.

..... 4. Dirty British coaster, with a salt-caked smokestack,
Butting through the Channel in the mad March days.

..... 5. In every adversity of fortune, to have been happy is the unhappy kind of misfortune.

..... 6. As a white candle
In a holy place,
So is the beauty
Of an aged face.

..... 7. The moan of doves in immemorial elms
And murmuring of innumerable bees.

..... 8. Saw a gloomy-gladed hollow slowly sink
To westward—in the deeps whereof a mere
Hound as the red eye of an eagle owl
Under the half-dead sunset glared.

..... 9. The chill
November dawn, and dewy glooming downs
The gentle showers, the smell of dying leaves,
And the low moan of leaden colored seas.

.....10. A good book is the precious life blood of a master spirit, embalmed and treasured on purpose to a life beyond.

TEST III: IDENTIFICATION TEST FOR ERRORS IN PHRASING OF THOUGHT

Among the many forms of errors in logical or in tasteful expression of thought are:

A. Anti-climax
B. Mixed metaphor
C. Inappropriate figure
D. Ponderous diction
E. Tautology
F. Redundancy
G. Over-elaboration

Place the letter for the error at the left of the first line of the example.

The pupil is advised to test each passage for the preceding errors and by the Method of Residues to eliminate all possibilities save the one finally chosen.

..... 1. With haggard eyes the Poet stood;
Loose his beard and hoary hair
Streamed like a meteor to the troubled air.

..... 2. By her own internal schism, by the abominable spectacle of a double Pope, the Church was rehearsing, as in still earlier forms she had already rehearsed, those vast rents in her foundation which no man should ever heal.

..... 3. As if the flower
That blows a globe of after-arrowlets
Ten-thousand fold had grown, flashed the fierce shield,
All sun.

..... 4. Then Nature tries the earth if it be in tune
And over it softly her warm ear lays.

..... 5. The Scripture moveth us to confess and acknowledge our manifold sins and wickedness.

..... 6. I will not suffer mine eyes to sleep,
Nor mine eyelids to slumber.

..... 7. The last of men was Dr. Johnson to have abetted squandering the delicacy of integrity by multiplying the labors of talents.

..... 8. There was no light in heaven save a few stars;
The boat put off o'ercrowded with their crews;
She gave a heel, then lurched to port
And going down, head foremost,—sunk, in short.

..... 9. She plows the billows like a hurricane, she throws the water from her bows, the wheels turn, the vessel starts.

.....10. Thy hair is as a flock of goats
That appeared from Mt. Gilead;
Thy teeth are like a flock of sheep
That are even shorn,
Which come up from the washing.

.....11. With nectar pure his oozy locks he laves.

.....12. In the effort to eradicate the scourge of intemperance the slumbering fires of passion were kindled.

.....13. The dawn is overcast; the morning lowers and heavily in clouds brings on the day.

.....14. That time of year thou may'st in me behold
When yellow leaves, or none, or few do hang

Upon those boughs which shake against the cold,
Bare ruined choirs where late the sweet birds sang.

.....15. From the silence and deep peace of this saintly summer night, from the pathetic blending of this sweet moonlight, dawnlight, dreamlight, suddenly as from the woods and fields, suddenly as from the chambers of the air opening in revelation, suddenly as from the ground opening at her feet, leaped upon her Death, the crowned phantom.

Test IV: Identification Test for Historical Classification of a Passage

English literature between 1700 and the present day has passed through four Periods, which may be termed—

A. Eighteenth century—classical period
B. Transition period
C. Nineteenth century—romantic period
D. Modern and free verse

A. Among the characteristics of eighteenth-century classical writing are critical analytic attitude, personifications of abstract ideas, heroic couplet, end-stopped lines, didactic morality, balance, antithesis, formal vocabulary, generalizations.

B. Transition literature combined Eighteenth-century style with Nineteenth-century thought and feeling.

C. Among the characteristics of Nineteenth-century romantic literature are sympathetic attitude especially to the oppressed; variety of figures, meters, rimes, rhythms; enthusiasm for nature; joy in the senses; effort toward beauty in thought and expression; personality of author revealed; mood of melancholy.

D. Among the characteristics of modern and free verse are avoidance of fixed rhythm, meter, or riming system; experimentation; grouping of non-related images; definite unusual images; unrestricted choice of subject; language of common speech.

The student should test the qualities of each selection, and identify it with one of the four historical periods.

Place the letter for the period at the left of the first line of the passage.

The student is advised, in case he recognizes the source of the quotation, to consider the selection here given only.

..... 1. Oh the wild joys of living! the leaping from rock to rock,
The strong rending of boughs from the fir tree, the cool silver shock
Of a plunge in a pool's living waters.

..... 2. Ill fares the land to hastening ills a prey
Where wealth accumulates and men decay.

..... 3. Ye who listen with credulity to the whispers of fancy and pursue with eagerness the phantoms of hope, who expect that age will perform the promises of youth, and that the deficiencies of the present day will be supplied by the morrow, attend.

..... 4. Sick for home
She stood in tears amid the alien corn.

..... 5. It's a warm wind, the west wind, full of birds' cries,
I never hear the west wind but tears are in my eyes;
It's a fine land, the west land, for hearts as tired as mine;
Apple orchards blossom there, and the air's like wine.

..... 6. Once more the ass did lengthen out
The hard dry hee-haw of his horrible bray.

..... 7. O wind, rend open the heat,
cut apart the heat, rend it sideways.
Fruit cannot drop through the thick air:
that presses up and blunts the points of pears
and rounds the grapes.

..... 8. He gave to Misery all he had,—a tear;
He gained from Heaven ('twas all he wished), a friend.

EXAMINATION VI

Examination VI is offered as a further example of objective measurement in a somewhat difficult field, viz., musical accomplishment.[1]

The fact that Examination VI is a standardized test does not lessen its value for our purposes. In fact, the teacher wishing to improve her objective-test methods can learn much from a critical study of various standard tests.

The teacher of music will note that this test rises above the level of the measurement of factual knowledge. It calls for a great deal of *performance* and *application* of technical information. Test 10, *Recognition of Familiar Melodies from Notation*, is particularly suggestive.

[1]The *Kwalwasser-Ruch Test of Musical Accomplishment*. Published and distributed by the Extension Division, State University of Iowa, Iowa City, Iowa. Reprinted by permission. Copyright, 1924, by Jacob Kwalwasser and G. M. Ruch.

KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT

For Grades IV-XII

Do not open this paper, or turn it over, until you are told to do so. Fill these blanks, giving your name, age, birthday, etc. Write plainly.

Name.....Date.....
(First name, initial, and last name)

Age last birthday.....years. Birthday.....
(Month and day)

Grade.....Teacher.....

School.....City.....

How many years have you studied music in school?.....

How long have you studied music outside of school?.....
(State your answer in half-hour lessons.)

Do not write below this line.

Test	Name of Test	Score
1	Knowledge of Musical Symbols and Terms	
2	Recognition of Syllable Names	
3	Detection of Pitch Errors in a Familiar Melody	
4	Detection of Time Errors in a Familiar Melody	
5	Recognition of Pitch Names	
6	Knowledge of Time Signatures	
7	Knowledge of Key Signatures	
8	Knowledge of Note Values	
9	Knowledge of Rest Values	
10	Recognition of Familiar Melodies from Notation	
Total		

Do Not Turn Over The Page Until The Signal is Given!

TEST 1

Knowledge of Musical Symbols and Terms

Directions: Below are twenty-five questions about music. Five answers are given to each question. Read each question and then draw a line under the right answer. The sample is already marked as it should be.

Sample: is called a **sharp natural flat note rest**

Begin here:

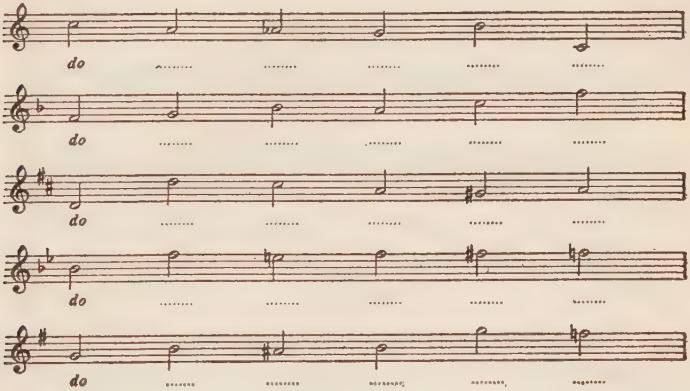
1	The first tone of the scale is	**mi re do fa sol**	1
2		is called a **rest natural sharp note flat**	2
3	The fifth tone of a scale is	**do fa mi sol re**	3
4		is a **flat note natural rest sharp**	4
5		is a **sharp flat natural note rest**	5
6		is a **slur hold rest double-sharp repeat-bar**	6
7		is called a **sharp flat natural note rest**	7
8		means **soft loud slow fast smooth**	8
9		is called a **bar staff measure accent clef**	9
10		is a **sharp flat natural note rest**	10
11		is a **clef staff measure accent phrase**	11
12		is called a **clef staff measure accent bar**	12
13		is a **clef measure staff phrase accent**	13
14		the curved line is a **slur tie hold accent rest**	14
15		is a **rest slur hold double-sharp repeat**	15
16		the curved line is a **slur hold rest tie accent**	16
17		means **higher lower louder repeat pause**	17
18		means **higher lower louder softer pause**	18
19	*Allegro*	means **lively slow repeat accent sweetly**	19
20	*f*	means **fast loud slow soft smooth**	20
21	*cresc.*	means **softer louder slower faster smooth**	21
22	*dim.*	means **smoother louder softer faster slower**	22
23	*Lento*	means **repeat accent sweetly slow lively**	23
24	*Legato*	means **soft quick separated connected loud**	24
25	*Staccato*	means **quick soft separated connected loud**	25

Test 1. Number right = Score.....

TEST 2

Recognition of Syllable Names

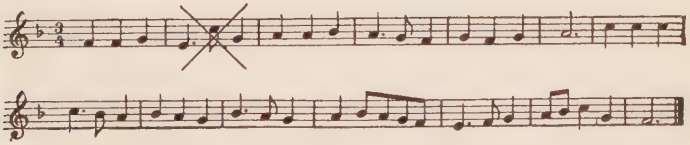
Directions: Below are five lines of notes. The first syllable in each line is **"Do"**; so the name do has been written below it. You are to write the syllable names on the lines under the other notes.

Begin here:

Test 2. Number right = Score.....

TEST 3

Detection of Pitch Errors in a Familiar Melody

Directions: The song "America" is written below. One measure has been crossed out because the melody is wrong. Five other measures are wrong. Hum over the melody to yourself and cross out all five wrong measures.

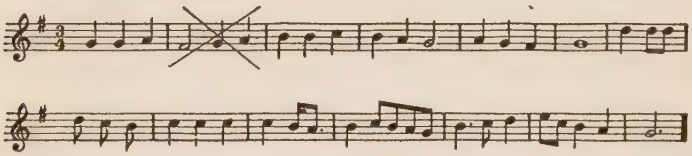
Begin here:

Test 3. Number right..... ×5 = Score.....

TEST 4

Recognition of Time Errors in a Familiar Melody

Directions: The song "America" is written below. One of the measures has been crossed out because it has the wrong number of beats. Five other measures are wrong. Hum over the song and cross out all five wrong measures.

Begin here:



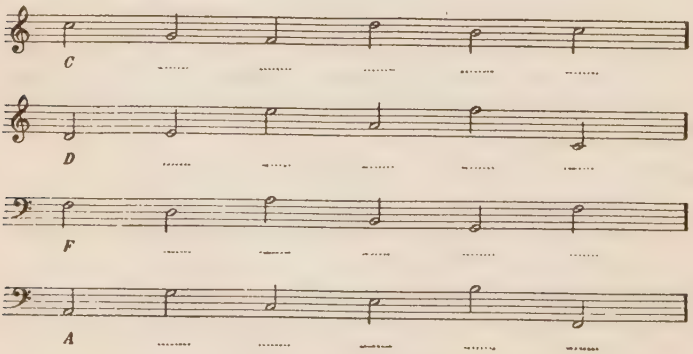
Test 4. Number right..... ×3 = Score.....

TEST 5

Recognition of Pitch Names

Directions: Below are four lines of notes. The first note in each line is already marked as it should be. You are to write the pitch or letter names on the lines under the other notes.

Begin here:

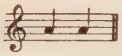


Test 5. Number right = Score.....

TEST 6

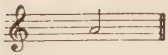
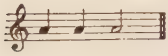
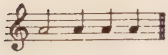
Knowledge of Time Signatures

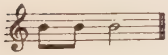
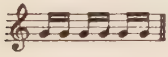
Directions: Below are ten full measures. At the right of each are five time signatures. You are to draw a line under the correct time signature for each measure. The sample is marked as it should be.

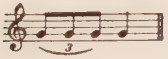
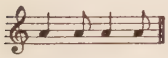
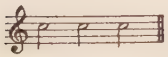
Sample:  The time signature is $\underline{\frac{2}{4}}$ $\frac{3}{4}$ $\frac{4}{4}$ $\frac{6}{8}$ $\frac{3}{8}$

Begin here:

1 The time signature is $\frac{2}{4}$ $\frac{3}{4}$ $\frac{4}{4}$ $\frac{3}{8}$ $\frac{9}{8}$ 1

2 The time signature is $\frac{2}{4}$ $\frac{3}{4}$ $\frac{4}{4}$ $\frac{6}{8}$ $\frac{9}{8}$ 2

3 The time signature is $\frac{3}{4}$ $\frac{4}{4}$ $\frac{6}{8}$ $\frac{9}{8}$ $\frac{3}{8}$ 3

4  The time signature is $\frac{6}{8}$ $\frac{4}{4}$ $\frac{5}{4}$ $\frac{3}{8}$ $\frac{2}{4}$ 4

5 The time signature is $\frac{2}{4}$ $\frac{5}{4}$ $\frac{4}{4}$ $\frac{3}{8}$ $\frac{3}{4}$ 5

6  The time signature is $\frac{3}{8}$ $\frac{2}{4}$ $\frac{4}{4}$ $\frac{3}{4}$ $\frac{6}{8}$ 6

7 The time signature is $\frac{5}{4}$ $\frac{4}{4}$ $\frac{2}{4}$ $\frac{3}{4}$ $\frac{6}{8}$ 7

8 The time signature is $\frac{3}{8}$ $\frac{9}{8}$ $\frac{2}{4}$ $\frac{6}{8}$ $\frac{4}{4}$ 8

9 The time signature is $\frac{2}{4}$ $\frac{3}{2}$ $\frac{4}{4}$ $\frac{6}{8}$ $\frac{3}{8}$ 9

10 The time signature is $\frac{2}{4}$ $\frac{6}{8}$ $\frac{9}{8}$ $\frac{3}{4}$ $\frac{4}{4}$ 10

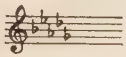
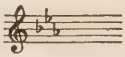
Test 6. Number right.....×2=Score.....

TEST 7

KNOWLEDGE OF KEY SIGNATURES

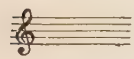
DIRECTIONS: At the left below is a column of ten major key signatures. At the right is a column of five minor key signatures. You are to write the names of the keys on the lines at the right of each signature.

Notice that there are two columns, one for major keys and one for minor.

SAMPLES:  D flat ____  C minor ____

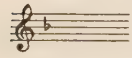
Begin here:

MAJOR KEY SIGNATURES

1  1

2 2

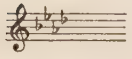
3 3

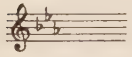
4  4

5 5

6 6

7 7

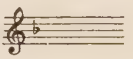
8  8

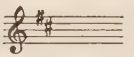
9  9

10 10

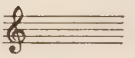
MINOR KEY SIGNATURES

11 11

12  12

13  13

14 14

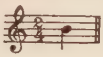
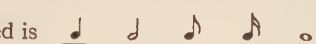
15  15

Test 7. Number right..... ×2 = Score.....

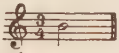
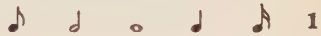
TEST 8

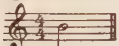
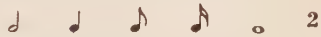
Knowledge of Note Values

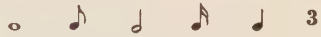
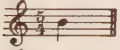
Directions: In the measures below a note has been left out of each. You are to draw a line under the note needed to complete the measure. The sample is already marked as it should be.

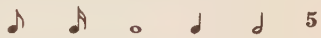
Sample:  The note needed is 

Begin here:

1  The note needed is  1

2  The note needed is  2

3 The note needed is  3

4 The note needed is 4

5 The note needed is  5

Test 8. Number right.....×3=Score.....

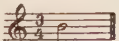
TEST 9

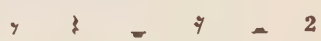
Knowledge of Rest Values

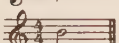
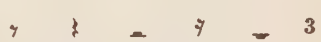
Directions: The five measures below are incomplete and need a rest to complete them. You are to draw a line under the rest needed to complete the measure. The sample is already marked as it should be.

Sample: The rest needed is

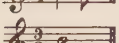
Begin here:

1  The rest needed is 1

2 The rest needed is  2

3  The rest needed is  3

4 The rest needed is 4

5  The rest needed is 5

Test 9. Number right.....×3=Score.....

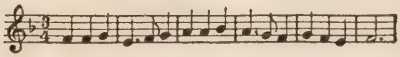
TEST 10

Recognition of Familiar Melodies from Notation

Directions: Below are phrases from ten songs that you know. Hum each line to yourself and then write the name of the song or the words of the phrase on the line at the right.

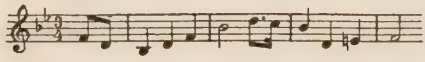
The sample is already marked as it should be.

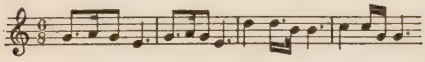
Sample:

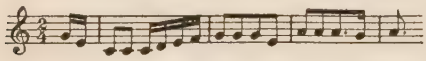


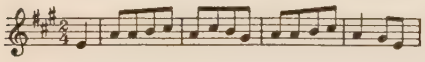
America or *My Country 'tis of Thee*

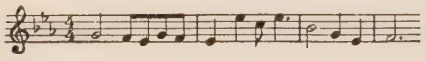
Begin here:

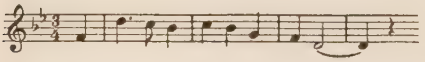
1  1

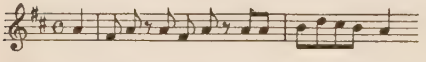
2  2

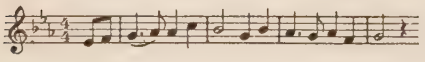
3  3

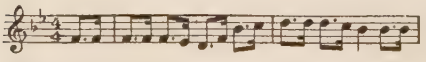
4  4

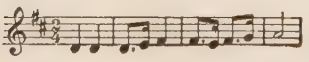
5  5

6  6

7  7

8  8

9  9

10  10

Test 10. Number right.....×5=Score.....

Addenda. Mention should be made of the growing tendency for writers of public-school textbooks and professional works on education to publish tests paralleling the content of such books. An early example is the series of tests to accompany Beard and Bagley, *The History of the American People*, the Macmillan Company. Scott, Foresman and Company publish a test paralleling Frasier and Armentrout, *An Introduction to Education.* The same publishers offer an extensive series of tests for use in connection with the Pieper-Beauchamp, *Everyday Problems in Science.* The Plymouth Press publishes a test by D. L. Geyer over the content of the *Twenty-first Yearbook of the National Society for the Study of Education*, which deals with intelligence testing. A number of other authors and publishers are now planning tests to accompany their textbook offerings.

Reference has already been made to the professional tests for selecting teachers in connection with the work of the Bureau of Public Personnel Administration. Many universities, particularly Columbia University, University of Iowa, University of Minnesota, and Ohio Wesleyan University, have prepared extensive tests for use with certain classes. Samples of these tests may be had in most cases. Particular mention should be made of the *Aptitude* and *Training* tests of the *Iowa Placement Examinations* series. These cover such college subjects as English, mathematics, physics, chemistry, modern foreign languages, law, etc., and are published by the University of Iowa Extension Division. For examples of some of the older Columbia examinations see B. D. Wood, *Measurement in Higher Education* (1923) World Book Company.

Weber has constructed a "Standard Achievement Test on Aims, Purposes, Objectives, Attributes, and Functions in Secondary Education"; publisher, Public School Publishing Company, Bloomington, Illinois. The tests prepared by the Summer Library Institute have already been listed.

For further samples of objective tests the reader is referred to the section on "Sample Tests" in the *General Bibliography* at the end of this volume and to the previously cited volume of Ruch and Rice, *Specimen Objective Examinations*.

In conclusion, the teacher is reminded that a careful study of the hundreds of standard tests in the elementary and high-school subjects will yield many invaluable suggestions as to approaches to the measurement of various school subjects. There are also a number of treatments of new-type or objective examinations listed in the section of the *General Bibliography* headed "Books, Monographs, and Bulletins." All of these contain sample examinations and tests.

For convenience the names of a number of leading publishers of standard tests are given below.

American Council on Education, 26 Jackson Place, Washington, D. C.
Chicago, University of, Press, 5750 Ellis Avenue, Chicago, Ill.
*Cincinnati, Bureau of Administrative Research of, University of Cincinnati, Cincinnati, Ohio
Courtis, S. A., 9110 Dwight Avenue, Detroit, Michigan
*Ginn and Company, 15 Ashburton Place, Boston, Mass.
Gregg Publishing Company, 20 West 47th St., New York City
Harvard University Press, Cambridge, Mass.
Houghton Mifflin Company, 2 Park Street, Boston, Mass.
*Iowa, University of, Extension Division, Iowa City, Iowa
Lippincott Company, East Washington Square, Philadelphia, Pa.
*Public School Publishing Company, Bloomington, Ill.
Scott, Foresman and Company, 623 South Wabash Avenue, Chicago, Ill.
Smith, Hammand and Company, Atlanta, Georgia
Southwestern Publishing Company, Cincinnati, Ohio
*Teachers College, Bureau of Publications of, Columbia University, New York City
*World Book Company, Yonkers-on-Hudson, New York
(Starred publishers are the larger distributors of standard tests.)

CHAPTER X

RULES FOR DRAFTING OBJECTIVE TEST ITEMS

TRUE-FALSE TESTS

Advantages and limitations. This chapter will deal with the principal forms of objective tests only. Their merits and demerits will be pointed out briefly, and certain rules for framing test items will be given. The true-false test will be considered first.

(*A*) Merits

1. Purely objective
2. Easy and rapid scoring
3. Can be made to measure reasoning as well as memory for facts
4. Rapidity of answering allows extensive sampling in limited time
5. Wide applicability

(*B*) Limitations

1. More difficult of construction than is commonly supposed (if ambiguities and partly-true-partly-false items are to be avoided)
2. Open to guessing and chance effects to a marked degree
3. Some subjects (e. g., the social and mental sciences) contain much that is controversial, and hence is neither absolutely true nor false.

Rules for constructing true-false items. These rules, or any rules for that matter, can be of little more than general assistance in framing true-false items. They do serve to

call attention to certain dangers and faults in test construction. The rules follow.[1]

1. *A true-false item (and any objective-test item) should observe the rules governing good language expression, grammar, spelling, punctuation, and capitalization.*

EXAMPLES

(*Poor*) Pasteurization *is where* milk is heated to about 165° F. for thirty minutes to retard souring. (Italicized statements here and in following examples show the faulty constructions, etc.)

(*Better*) Pasteurization is the heating of milk to about 165° F. for thirty minutes in order to retard souring.

(*Poor*) Sufficient data *is* at hand to suggest that speed and accuracy are closely related in arithmetic computation.

(*Better*) Sufficient data are at hand to suggest that speed and accuracy are closely related in arithmetic computation.

(*Poor*) Benedict Arnold was a *noted* traitor to his country.

(*Better*) Benedict Arnold was a notorious traitor to his country.

2. *Avoid the use of double negatives, since these increase the reading difficulty of the item and thus tend to throw success or failure upon a basis of reading comprehension rather than knowledge of subject-matter.*

EXAMPLES

(*Poor*) Freezing weather is *not entirely unknown* in Florida. (True)

(*Better*) Florida occasionally has freezing weather.

(*Poor*) Scientists believe that perpetual motion is *not impossible.* (False)

(*Better*) Scientists believe that perpetual motion is possible.

3. *Avoid introducing "trick," "catch," or puzzle questions.*

EXAMPLES

(*Poor*) The Civil War began in 1861 *B. C.* (False)

(*Better*) The Civil War began in 1861 *A. D.* (Or omit the *A. D.* entirely. If a false question is desired, change 1861 to some other date.)

(*Poor*) *Runyan* Kipling wrote "Just So Stories." (False) (NOTE: Scored "false" because of error in first name of author.)

(*Better*) Rudyard Kipling wrote "Just So Stories." (Or, if a false item is desired, substitute the name of some other author.)

[1] For additional rules, *see* Weidemann, *How to Construct the True-False Examination*, pp. 41-73. For similar rules for other types of objective tests, *see* Paterson, *The Preparation and Use of New-Type Examinations*, pp. 14-66.

4. *Avoid the use of items which are partly true and partly false.* Some testers have made wide use of such statements, instructing the pupils to "mark the statement false *if any part of it is false.*" Such statements also tend to fall to the level of "catch" questions. A possible exception may be made with advanced college classes.

EXAMPLES

(*Poor*) The battle of Gettysburg, which is usually considered to be the turning-point of the Civil War, was fought in *1862.* (False)

(*Better*) The battle of Gettysburg was fought in 1862. (Note: Since the falsity of the item rests upon the date alone, omit the clause dealing with the turning-point of the war.)

(*Poor*) Poe's writings are characterized by his bizarre imaginings, originality of theme, *faithful character portrayals,* and skill in arousing interest. (False)

(*Better*) Poe's writings are characterized by their fidelity to historic events, faithful character studies, and conventional vocabularies. (False)

(*Poor*) Many experts consider the income tax to be a very just form of taxation because such a tax *can* readily be passed on or shifted to another. (False)

(*Better*) Many experts consider the income tax to be a very just form of taxation because such a tax cannot readily be passed on to another. (True)

5. *Avoid long sentences with many dependent or modifying clauses.* There is no good reason why a true-false statement must be a single sentence. For the sake of easier comprehension, it is often better to form two or three shorter sentences. It should be remembered that the true-false test should measure knowledge of subject-matter, not skill in reading.

EXAMPLES

(*Poor*) The digestion of starches, which is commenced in the mouth by the action of the ptyalin of the salivary secretions, is stopped when the food reaches the stomach, which has an acid reaction due to the presence of hydrochloric acid; ptyalin being active only in neutral or alkaline media. (True)

(*Better*) The digestion of starches is begun in the mouth under the influence of the ptyalin of the salivary secretions. Ptyalin is active only in neutral or slightly alkaline media. For this reason the acid of the gastric juice soon stops the digestion of starches after the food reaches the stomach.

Even in the second phrasing of this item, care must be taken to avoid the situation of the partly-true-partly-false type of item. It would be better to break such an item into two or three separate items, somewhat as follows:

1. The digestion of starches is begun in the mouth under the influence of the enzyme, ptyalin. (True)
2. Ptyalin operates only in an acid medium. (False)
3. The gastric juices of the stomach are alkaline in reaction. (False)
Etc.

6. *Avoid words which prejudice pupils' replies.* Such words as "always" and "never" occur in false statements about *twice* as often as they occur in true statements, according to Weidemann. If this is true, pupils soon learn that such words offer "clues" to the right answers. Weidemann calls these *specific determiners.*

EXAMPLES

(*Poor*) Animals *always* have the power of locomotion. (False)

(*Better*) Some animals cannot move from place to place unless carried by outside forces. (True)

A pupil ignorant of the fact that certain animals are sessile, will mark the first statement "false" simply because the word "always" appears in the statement. He might be quite ignorant of the fact that certain animal forms are incapable of independent movement.

(*Poor*) Water *never* runs up hill. (False)

A pupil may be quite uninformed of the facts about siphons, pumps, etc., where water rises for short distances, and yet answer "false" because of the word *never.*

7. *Choose simple, everyday words in preference to more technical or literary synonyms in framing true-false items.*

EXAMPLES

(*Poor*) California and Florida produce large quantities of *citrus* fruits. (True)

(*Better*) California and Florida produce large quantities of oranges, lemons, and grapefruit.

(*Poor*) Good harbors are often formed by diastrophism. (True)
(*Better*) Good harbors are often formed by the gradual sinking of coast lines near the mouths of rivers.

8. *Avoid having two items in the same test (or at least near together) the answer to one of which suggests the answer to the other.*

Item 1. It is thought that the Norsemen visited America earlier than the first voyage of Columbus. (True)
Item 5. No white man saw the shores of America before the time of Columbus. (False)

Items 1 and 5 are likely to prove mutually helpful in answering *both* items; one should probably be omitted.

SIMPLE-RECALL TESTS

Advantages and limitations. Simple-recall questions form one of the most widely used objective test types. If carefully phrased, such tests form an almost ideal compromise between the completion type proper and the recognition types (true-false, multiple-choice, etc.); the former being not quite objective, and the latter being objective but open to guessing.

The simple-recall test is almost perfectly objective, and it is little subject to chance effects.

(*A*) Advantages
1. Almost entirely objective
2. Guessing and chance scores almost negligible
3. Fairly easy and rapid scoring
4. A "natural" form of questioning; similar to the usual oral or written question

(*B*) Limitations
1. Not quite perfectly objective
2. Tend to be factual in character (as a matter of avoiding subjectivity in scoring)
3. Scoring somewhat more laborious than in the case of recognition types

Rules for constructing simple-recall items. A few general comments are in order:

1. *Avoid items which can be answered by the exercise of general intelligence without knowledge of the subject-matter concerned.*

EXAMPLES

(*Poor*) The force of gravitation causes water to run *down hill*.

(*Better*) The force which causes water to run down hill is called *gravitation*.

(*Poor*) In the autumn, deciduous trees shed their *leaves*.

(*Better*) Trees which shed their leaves in the autumn are termed *deciduous*.

2. *Since several answers for a given blank may be expected at times, make all meritorious answers center about a single idea.* If all possible answers which are worthy of credit are really nothing but variations of verbal statements of a single idea, little subjectivity results.

EXAMPLES

(*Poor*) Fruit may be kept for long periods by *refrigeration*.

Other possible and meritorious answers are: "wrapping," "drying," "canning," "freezing," etc.; or even "dealers," "merchants," "people," etc. Such an item is of little value.

(*Better*) Fresh fruits are often kept from spoiling during long shipments by.....

"Refrigeration," "ice," "icing," "cooling," etc., are almost certain to arise as answers, but these are close synonyms.

3. *Make the response call for a single word, date, short phrase, or in general a very short response.*

(*Poor*) The Civil War started over.....

The expected reply was "secession." However, there are many very different answers of considerable or equal merit which are certain to arise, e. g., "the firing on Fort Sumpter," "states' rights," "slavery," etc.

(*Better*) An immediate cause of the outbreak of the Civil War was the withdrawal from the Union of the State of *South Carolina*.

4. *Avoid using the words "a" and "an" immediately before a blank whenever possible.* These words offer slight clues to the expected answers. Possible solutions are (*a*) re-wording to avoid the indefinite article or (*b*) the use of such a device as "a(n)." Method (*b*) is somewhat unnatural but adequate.

EXAMPLES

(*Poor*) The unit of measurement of electric resistance is an.....

(*Better*) The unit of measurement of electric resistance is the.....

.....

(*Better*) The unit of measurement of electric resistance is a(n).....

.....

In the above case the only very likely answers are ohm, volt, ampere, or watt. Either "a" or "an" in such items reduce the probable field of choice to two possibilities, instead of four.

5. *Terminal and aligned blanks are more convenient in scoring than are staggered blanks.*

EXAMPLES

(*Poor*) .....wrote "The Raven."

There are.....lines in a sonnet.

The men of Odysseus were changed into.....by Circe.

(*Better*) "The Raven" was written by.....

The number of lines in a sonnet is.....

Circe changed the men of Odysseus into.....

COMPLETION TESTS

Advantages and limitations. It should be remembered that the completion test originated as a test of general intelligence at the hands of Ebbinghaus. Its use as a subject-matter test requires that it be greatly modified from the form in which Ebbinghaus used it.

(*A*) Advantages

1. Very free from guessing and chance effects
2. May be used in almost any school subject
3. Allows some freedom of expression and reasoning
4. Is a "natural" form of questioning as it parallels the thought processes
5. Is easy to prepare

(*B*) Limitations

1. Not highly objective unless great care is taken
2. Unless the number of blanks in a given passage is kept small, the completion test tends to call for the exercise of general intelligence to too large an extent
3. Difficult to score because of the staggered arrangement unless cut-out stencils are used

Rules for constructing completion tests. These are naturally similar to those for simple-recall items in many respects.

1. *Make each blank call for a single idea.* (See Rule 2 for simple-recall tests.) It does not introduce marked subjectivity if there are numerous equivalent answers, provided these are listed on the scoring key or stencil.

2. *Avoid a large number of blanks in a single sentence or paragraph. Omit only a few critical words.*

EXAMPLES

(*Poor*) The walls of the stomach secrete the *gastric juice* which acts on *proteins*, changing them into *peptones* and *proteases*. The active enzyme is *pepsin* but the gastric juice also contains an *acid* which activates the *pepsin*.

(*Better*) The gastric juice contains the enzyme *pepsin* which acts on *proteins*. It also contains *acid* which makes the pepsin active.

(*Poor*) The principal *battle* of the *Civil* War was *Gettysburg*.

(*Better*) The battle of *Gettysburg* is usually considered to be the turning-point of the Civil War.

The first statement of the history item did not make clear whether the first blanks referred to battles, causes, generals, results, or what, nor is it clear what war is meant. The second form is better, although less broad in scope.

3. *Make all blanks of the same length in order to avoid giving clues to the length of the answer expected.*

EXAMPLES

(*Poor*) The process of food manufacture in green plants is called *photosynthesis*. The raw products of this process are *water* and *carbon-dioxide*.

(*Better*) The process of food manufacture in green plants is called..... ..... The raw products of this process are..... and.....

4. *Do not attempt (as is sometimes done) to make each dot of a blank stand for a letter of the required word.* Some test workers use this plan rather largely. It is open to the objection that a pupil may know the answer but will not hit upon the particular synonym that the test maker had in mind when he drafted the item.

EXAMPLES

(*Poor*) A precious metal heavier than lead is .....

The expected answer was "platinum" but "gold" or "iridium" are equally good answers.

(*Better*) A precious metal heavier than lead is .....

(*Poor*) Plants with scattered fibro-vascular bundles are called .....

The thirteen dots were intended to suggest "monocotyledons," although "endogens" is equally good as an answer.

5. *It is ordinarily unwise and uneconomical to make completion tests by deleting occasional words from an actual passage in the textbook.* Better completion exercises may be obtained by actual writing of suitable sentences or paragraphs de novo.

For obvious reasons, it is difficult to give examples here. The teacher can test the soundness of this advice by actual

attempts to make completion exercises from quotations from a textbook. It can be done at times, although the procedure is wasteful.

MULTIPLE-RESPONSE TESTS

Advantages and limitations. Although the true-false test is, in one sense, a multiple-response test, we shall treat the latter here, as elsewhere, as a separate variety of *recognition* types. It is often convenient to view the simple-completion (recall) and the completion test proper as *recall* tests in the generic sense that they suggest little or nothing about the expected answer. Multiple-choice, true-false, matching, rearrangement, etc., types fall under the generic heading of *recognition* tests. Even the two-response test, which has some features in common with true-false tests (e.g., 50:50 chance of guessing correctly; at least in theory), will be shown in a later section to differ importantly from the true-false type of item. The rules governing the construction of multiple-choice and true-false tests are dissimilar enough to justify separate treatments.

(*A*) Merits

1. Fairly easy to construct
2. Purely objective
3. Usually more reliable than true-false tests, but ordinarily not quite so reliable as well-constructed simple-recall tests (when equal numbers of items are considered)
4. May be made to test reasoning as well as facts if the response items are made long statements. The reference here is to the *best-answer* type of multiple-choice items. The best-answer test lends itself to the use of several more or less extended statements which may be made to vary in merit from *one* entirely acceptable statement, through gradual degrees to one or more statements which are quite lacking in merit.

5. A sufficient number of statements can be used to minimize guessing to any desired degree (within practical limits).

(*B*) Limitations

1. Test makers must guard against allowing these tests to become purely fact items.

2. They are likely to be space-consuming, especially if the best-answer variety is employed.

3. It is often difficult to find from three to seven responses which present reasonable plausibility. In a four-response test, for example, if two responses are obviously unrelated and absurd, the choice often reduces to a situation of but two responses.

Rules for constructing multiple-response tests. The following suggestions may prove helpful.

1. *Use at least four or five responses whenever possible in order to minimize chance successes.*

2. *Choose the responses so that all, or at least most, of them have some degree of plausibility.*

EXAMPLES

(*Poor*) The first president of the United States was John Adams *Henry Ford Jack Dempsey* George Washington *Douglas Fairbanks*

(*Better*) The first president of the United States was John Adams Thomas Jefferson George Washington James Madison Andrew Jackson

3. *Avoid wordings which serve as clues, e.g., changes in parts of speech, mixed singular and plural responses, etc.*

EXAMPLES

(*Poor*) "Candid" means incisive *candy* frank *slavery* cowardly

(*Better*) "Candid" means incisive secretive frank abrupt crafty

(*Poor*) The name for the result in division is sum *factors* quotient *remainders products*

(*Better*) The name for the result in division is sum factor quotient remainder product

4. *Avoid, when possible, the use of "a" or "an" as the final word prior to the listing of the responses; these words act as hints or clues.*

(*Poor*) The starfish is an insect sponge echinoderm protozoan coelenterate

(*Better*) The starfish belongs to the group of insects sponges echinoderms protozoans coelenterates

(*Better*) The starfish is a(n) insect sponge echinoderm protozoan coelenterate

5. *Make the first, second, third, etc., responses the correct response in about equal numbers.*

6. *Do not mix items with varying numbers of responses in the same test if the scores are to be corrected for chance* [by the formula, $\text{Score} = R - W/(n-1)$].

7. *Do not allow the correct response to occur in the same position (order) for more than two or three successive items.*

MATCHING TESTS

Advantages and limitations. Matching tests have certain features not common to the other types of objective tests.

(*A*) Merits

1. Purely objective
2. Easily constructed for certain types of subject matter
3. Rapidly scorable
4. May be used to measure either factual mastery or judgment
5. Chance successes may be avoided by using ten or more pairs, or incomplete matchings.

(*B*) Limitations

1. Much subject-matter does not lend itself to this method.
2. If five or fewer pairs are used, chance enters appreciably in success or failure.
3. Very long exercises (25-30 or more pairs) are wasteful of pupils' time in searching out the proper pairings (without

bringing adequate compensations by way of reducing guessing).

4. When matching exercises deal with dates and chronologies, the use of large numbers of pairs tends to throw dates so close together that such fine discriminations cannot be justified upon the grounds of social utility.

Rules for constructing matching tests. The following rules are based principally upon experimental findings of the author and his students.

1. *The optimum number of pairs to be matched is probably between ten pairs and twenty pairs; fewer than ten introduces considerable of the chance element, and more than twenty is decidedly wasteful of time.*

2. *If fewer than ten complete pairs are to be matched, make an excess of statements in one column or the other.*

EXAMPLES

1453	(1534)	Cartier discovered the St. Lawrence River
1492	( )	Sea trade with India established by da Gama
1498	( )	Defeat of the Spanish Armada
1534	( )	Capture of Constantinople by the Turks
1588	( )	First voyage of Columbus
1608		
1619		
1754		

3. *Avoid such clues as having some words plural and some singular.*

EXAMPLE

1. *Enzymes*	A digestive gland
2. *Radius and ulna*	A large artery
3. Pancreas	*Bones of the arm*
4. Aorta	The wind-pipe
5. Trachea	*Digestive agents*

4. *Avoid having a small number of dates, or other distinctive facts, in a general list, since these obviously reduce the field of choice to selection among a very few alternatives.*

EXAMPLE

1. *Battle of Bunker Hill*	....3....	Andrew Jackson
2. Louisiana Purchase	.....	Herbert Hoover
3. Spoils System	.....	Goethals
4. Food relief in Belgium	.....	*1775*
5. *Famous panic*	.....	Thomas Jefferson
6. Forest conservation	.....	*1865*
7. **Eighteenth Amendment**	.....	Pinchot
8. *Lee's surrender*	.....	**Prohibition**
9. Panama Canal	.....	**Slavery**
10. **Thirteenth Amendment**	.....	*1893*

In the above example the pairing of the italicized items and responses (or the bold-face items and responses) is far more likely, by guess, than seems to be true at first sight.

Part III

EXPERIMENTAL AND THEORETICAL CONSIDERATIONS

CHAPTER XI

EXPERIMENTAL STUDIES ON NEW-TYPE EXAMINATIONS[1]

Introduction. This chapter will summarize the principal comparative, experimental studies of the merits of the various types of objective test items under four general headings:

I. Studies of Comparative Validities
II. Studies of Comparative Reliabilities
III. Studies of Comparative Working Times
IV. Studies of Comparative Difficulties

COMPARATIVE VALIDITIES

Brinkley's study. Brinkley[2] compared a number of types of tests by correlating each against a general criterion made up of a large number of tests (both old- and new-type), class marks, teachers' judgments, and pupils' judgments. These were combined into a single measure. Each objective test was correlated against this criterion. He also made two sub-criteria, one for *thought* and the other for *information* values of the tests and examinations. Selected findings by Brinkley are summarized in Tables 25 and 26.

Brinkley's results show the objective types of tests to be somewhat superior to the essay examinations. The com-

[1] Part III (especially chapters XI, XII, and XIII) is intended primarily for serious students of educational measurement. It is to be hoped, however, that the classroom teacher will be interested in the more critical phases of examination construction discussed in Part III. If the teacher's grasp of the theory and practice of educational measurement is to rise above rule-of-thumb methods, certain experimental and statistical results must be examined critically.

[2] S. G. Brinkley, "Values of the New-Type Examinations in the High School," *Teachers College Contributions to Educations*, No. 161 (New York: Columbia University, 1924), especially pages 68-82 and 84-90.

TABLE 25

BRINKLEY'S CORRELATIONS OF 31-MINUTE TESTS WITH HIS GENERAL CRITERION*

TYPE OF TEST	CORRELATION WITH GENERAL CRITERION
True-false	.82 ± .023
Multiple-choice	.82 ± .023
Completion	.84 ± .021
Word or phrase answer	.86 ± .018
Rearrangement	.82 ± .023
Essay Test I	.81 ± .024
Essay Test II	.76 ± .029
Fall term marks	.65 ± .040
Teachers' judgments	.88 ± .016
Pupils' judgments	.83 ± .022
Otis Intelligence Test scores	.55 ± .048

*Quoted from Brinkley's Table V, *op. cit.*, p. 85.

pletion and word-or-phrase-answer types seem to be somewhat superior to the true-false and multiple-choice varieties, but these differences are too small to be statistically significant.

Table 26 below reports the correlations of the seven different types of tests with Brinkley's thought and information criteria.

TABLE 26

BRINKLEY'S CORRELATIONS OF 31-MINUTE TESTS WITH HIS INFORMATION AND THOUGHT CRITERIA*

TYPE OF TEST	CORRELATION WITH INFORMATION CRITERION	CORRELATION WITH THOUGHT CRITERION
True-false	.75†	.70†
Multiple-choice	.76	.64
Completion	.78	.74
Word or phrase answer	.76	.75
Rearrangement	.70	.64
Essay (2 tests)	.73	.73
Intelligence	.66	.70

*Quoted from Brinkley's Table VII, *op. cit.*, p. 89.
†The probable errors range from .031 to .041.

Although the correlation of the new-type tests with the thought criterion are somewhat lower than with the information criterion, the general sweep of the evidence supports the conclusion that the new-type tests are at least as valid as the essay examinations; especially the completion and word-or-phrase-answer tests.

The experiments of DeGraff and Ruch. DeGraff and Ruch, working under a subvention from the New York Commonwealth Fund, studied a number of types of objective tests. As a criterion, two simple-recall tests were constructed so as to be *equivalent forms*. The subject-matter was that of United States history. A total of 2533 pupils took both of these recall tests. Each pupil then took the "same" items once more on a different day in some recognition form (true-false, two-response, three-response, five-response, or seven-response). The groups taking the recognition forms were further subdivided by the plan of having half take the test with instructions *to guess* (when in doubt) and half with directions *not to guess*.

To summarize, all pupils first took Recall, Form A (100 items) and Recall, Form B (100 items); the total group was then divided by chance into ten sub-groups, each sub-group taking one of the five above mentioned recognition editions of the "same" items with instructions either for or against guessing.

The word "same" has been placed in quotation marks because it is doubtful whether the items may be held to be the same when changed from recall to the various recognition types. The degree to which the items remained the same throughout the various editions of the tests may be judged by the following sample item.

Recall: Eli Whitney is noted for his invention of the.....

7-response: Eli Whitney is noted for his invention of the (1) steamboat (2) spinning jenny (3) cotton gin (4) telegraph (5) telephone (6) printing press (7) steam engine.

5-response: Eli Whitney is noted for his invention of the (1) steamboat (2) spinning jenny (3) cotton gin (4) telegraph (5) telephone.

3-response: Eli Whitney is noted for his invention of the (1) spinning jenny (2) cotton gin (3) telegraph.

2-response: Eli Whitney is noted for his invention of the (1) spinning jenny (2) cotton gin.

True-false: Eli Whitney is noted for his invention of the spinning jenny.

Table 27 shows the validity coefficients (correlations of each test against the recall test as a criterion).[1]

TABLE 27

INTERCORRELATIONS, CORRECTED AND UNCORRECTED FOR CHANCE, FOR ALL TEN TESTS USED

TYPE OF TEST	RECALL A VS. RECOGNITION A		RECALL B VS. RECOGNITION B	
	Uncorrected	Corrected	Uncorrected	Corrected
7-response (g)*	.871 ± .011	.873 ± .011	.816 ± .015	.861 ± .111
7-response (n)†	.927 ± .006	.926 ± .006	.872 ± .012	.898 ± .009
5-response (g)	.907 ± .008	.910 ± .008	.860 ± .011	.903 ± .008
5-response (n)	.891 ± .009	.918 ± .007	.836 ± .013	.870 ± .010
3-response (g)	.838 ± .013	.848 ± .012	.797 ± .016	.875 ± .010
3-response (n)	.845 ± .014	.915 ± .007	.852 ± .012	.902 ± .008
2-response (g)	.859 ± .012	.865 ± .011	.735 ± .021	.806 ± .016
2-response (n)	.740 ± .018	.775 ± .016	.752 ± .018	.868 ± .010
True-false (g)	.804 ± .015	.839 ± .013	.675 ± .024	.801 ± .016
True-false (n)	.749 ± .018	.860 ± .011	.768 ± .017	.856 ± .011
Correlation of Recall A vs. Recall B				.950 ± .001
Coefficient of reliability (sum of Recall A and B)				.974 ± .001

*(g) indicates the tests taken under instructions to guess.
†(n) indicates the tests taken under instructions not to guess.

Table 27 shows clearly that the recognition types of tests (if the recall tests may be accepted as a valid criterion)

[1]Abridged from G. M. Ruch *et al, Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Co., 1926), p. 75. (The details of this study are given in this reference.) A briefer report is given in the *Journal of Educational Psychology*, Vol. XVII (September, 1926), pp. 368-375.

measure roughly the same abilities or functions. The correlations are moderately high in all cases, although it appears to be true that the larger the number of responses, per item, the more valid the test. The correlations presented are never close to unity (1.00, or perfect correlation) for a number of reasons, particularly, (*a*) the fact that both the recall tests and the recognition tests are unreliable to some degree and hence cannot correlate perfectly, and (*b*) the fact that recall and recognition tests undoubtedly measure somewhat different abilities. *The rough indication is that the various types of tests studied are not greatly unequal in validity, although true-false and two-response tests are less valid than the tests with a larger number of optional responses.*

It is further true that *instructions against guessing combined with the use of the chance correction formula* $S = R - \frac{W}{n-1}$ *give somewhat more valid results than when directions to guess are employed* (whether corrected or uncorrected). This point will receive further attention in a later chapter.

It would be interesting to know whether the lack of perfect correlation between the criterion (recall tests) and various types of recognition tests is due principally to (*a*) unreliability or (*b*) differences in the abilities measured. The only way to attack this question is through recourse to what the statistician calls "correction for attenuation." Coefficients of correlation may be corrected for the dilution arising from unreliability by the use of certain formulas derived by Professor Spearman and others. The resulting *corrected coefficients of correlation* are to be thought of as estimates of the probable correlation which would be found *if perfectly reliable measures had been used.*[1]

[1]The form of the formula for corrections for attenuation used in obtaining the values in Table 28 is:

$$r_{\infty x \infty y} = \frac{\sqrt{r_{x_1y_1} \cdot r_{x_2y_2}}}{\sqrt{r_{x_1x_2} \cdot r_{y_1y_2}}}$$

Where:

$r_{\infty x \infty y}$ is the corrected coefficient of correlation
$r_{x_1y_1}$ is the correlation of Recall A with Recognition A
$r_{x_2y_2}$ is the correlation of Recall B with Recognition B
$r_{x_1x_2}$ is the correlation of Recall A with Recall B
$r_{y_1y_2}$ is the correlation of Recognition A with Recognition B

Table 28 shows the corrected coefficients. With two exceptions the uncorrected coefficients are above 0.900. When scores are corrected for chance, the values are always over 0.900. These results prove rather definitely that *recall and recognition types measure roughly the same abilities.*

TABLE 28*

CORRELATION OF RECALL AND RECOGNITION WHEN CORRECTED FOR ATTENUATION

GUESS	UNCORRECTED FOR CHANCE	CORRECTED FOR CHANCE
7-response	.967	.971
5-response	.974	.975
3-response	.916	.954
2-response	.945	.921
True-false	.943	.953
DO NOT GUESS	**UNCORRECTED FOR CHANCE**	**CORRECTED FOR CHANCE**
7-response	.980	.982
5-response	.953	.976
3-response	.925	.988
2-response	.838	.917
True-false	.827	.962

**Objective Examination Methods in the Social Studies*, p. 77

The reader unfamiliar with statistical methods may regard the corrected coefficients of correlation as promises of what would be obtained if infinitely long (and hence perfectly reliable) tests of both recall and recognition types had been correlated. It cannot well be maintained that true-false and multiple-choice tests do not measure roughly the same abilities as recall tests in the light of this evidence. Some writers and teachers have held that true-false tests, especially, were largely matters of chance and hence not trustworthy. Although true-false and two- and three-response tests are open to much guessing and chance successes or failures, nevertheless these tests, if made long enough to compensate for guessing, do measure roughly the same functions as do the recall tests whose validity is never seriously questioned.

There is one disturbing factor in Table 28 which needs comment, viz., the fact that this table compares various recognition tests of 100 items each, but overlooks the fact that the working times are very different for such tests and the true-false and seven-response. The question may well be asked whether comparisons upon the basis of *equal working times* rather than *equal numbers of items* would not be fairer. The answer is undoubtedly in the affirmative, and such comparisons will be made when the question of relative reliabilities is discussed.

Wood's investigations. Wood has done invaluable work in comparing the relative validities of old- and new-type examinations. In one report[1] he gives extensive data on true-false tests in a number of college subjects. Like the previously reported study of DeGraff and Ruch, validity coefficients are reported for both corrected and uncorrected (for chance) scores. Wood's criteria differ somewhat from one subject to the next; usually being a combination of other tests, instructors' marks, and old-type examinations. The data of Table 29 are selected from his report.

TABLE 29

SELECTED VALIDITIES REPORTED BY WOOD FOR COLLEGE COURSES

SUBJECT	NO. OF ITEMS	SCORE = NO. RIGHT	SCORE = RIGHTS MINUS WRONGS
French	100	.706	.747
Law (Pleading and Practice)	180	.845	.868
Law (Property)	200	.669	.709
Law (Torts)	130	.761	.815
Anatomy*	130	.654	.632
Anatomy†	130	.649	.640
Anatomy**	130	.766	.766
Averages	...	.721	.769

*Criterion here is average of three one-hour essay examinations.
†Criterion here is average of all first-year medical grades except anatomy.
**Criterion here is a 200-item completion test.

[1]B. D. Wood, "Studies of Achievement Tests," *Journal of Educational Psychology*, Vol. XVII (1926), pp. 1-22, 125-129, and 263-269.

Using as a criterion six essay examinations, Wood has calculated the validity coefficients given in Table 30 for the three law examinations.

TABLE 30

VALIDITIES OF THREE LAW EXAMINATIONS AS GIVEN BY WOOD

SUBJECT	NO. OF ITEMS	SCORE = NO. RIGHT	SCORE = RIGHTS MINUS WRONGS
Pleading and Practice	180	.688	.744
Property	200	.705	.745
Torts	130	.605	.674
Averages	...	.666	.721

Although only true-false tests are involved in this study by Wood, the results give confidence in the conclusion that true-false tests are measures of mastery of subject-matter. It should be pointed out that these law examinations are by no means tests of memory for facts, but that they also include a great deal of reasoning. The correlations reported are far from perfect, but this is not surprising as the criteria are also far from ideal measures.

The study by Paterson and Langlie. These authors[1] report findings on a 100-item true-false test in general psychology. The criterion of validity here is average scholarship. The validity coefficients are reported for both "rights" and "right-minus-wrong" scores.

METHOD OF SCORING	VALIDITY	NO. OF ITEMS	NO. OF CASES
No. Right	.44 ± .050	100	111
Rights minus Wrongs	.39 ± .054	100	111

These validity coefficients are very low in comparison with the studies reported earlier. Two reasons are suggested

[1]D. G. Paterson and T. A. Langlie, "Empirical Data on the Scoring of True-False Tests," *Journal of Applied Psychology*, Vol. IX (1925), pp. 339-348.

for such low values: (*a*) average scholarship is a very fallible criterion, and (*b*) the true-false test used was rather too easy for good discrimination (the average score being about 84 for "rights" and 71 for "rights-minus-wrongs." It is also possible that this examination was not very adequate; indeed, Paterson and Langlie report its reliability as 0.63 for "rights" and but 0.54 for "rights-minus-wrongs." Such values are certainly low for 100-item true-false tests as reported by other investigators. (See Charles's results, which follow.)

Charles's study of five types of objective items in psychology examinations. Charles has carried on under the author's direction a much more extensive study than that reported by Paterson and Langlie, and with rather different results.[1] The general plan of the investigation was that devised by DeGraff and Ruch. Table 31 gives a summary of Charles's findings.

TABLE 31

CHARLES'S INVESTIGATION OF THE VALIDITIES OF FIVE TYPES OF OBJECTIVE TESTS IN ELEMENTARY PSYCHOLOGY

TYPE OF TEST	NO. OF ITEMS	VALIDITY COEFFICIENTS			
		CRITERION = RECALL SCORES		CRITERION = TERM MARKS	
		Rights	R−W	Rights	R−W
5-Response.....	50	.71 ± .024	.71 ± .020	.21 ± .047	.23 ± .047
3-Response.....	50	.70 ± .024	.71 ± .023	.30 ± .030	.26 ± .046
2-Response.....	50	.64 ± .029	.66 ± .026	.32 ± .044	.41 ± .041
True-false	50	.68 ± .026	.70 ± .024	.23 ± .046	.26 ± .047
Recall.....	50	.....	.....	.31 ± .022	.....

Charles's validity coefficients computed against the criterion of term marks are even lower than those reported by Paterson and Langlie for average scholarship. Charles's

[1] J. W. Charles, *A Comparison of Five Types of Objective Tests in Elementary Psychology* (1926), unpublished doctor's dissertation, University of Iowa. A summary appeared in the *Journal of Applied Psychology*, Vol. XII (1928), pp. 398-403.

tests, however, were only half as long as those used by Paterson and Langlie. When the recall test scores are taken as a criterion, the validities are much higher; in fact, Charles showed that the correlations between recall and recognition types were almost as high as could be expected in view of the unreliabilities of the measures.[1] If allowance for unreliabilities is made, the recall and recognition tests may be said to measure substantially the same abilities.

Summary of the studies on comparative validities. The following conclusions seem justified in view of the findings of Brinkley, DeGraff and Ruch, Wood, Paterson and Langlie, and Charles:

1. Where old- and new-type tests are compared, the new-type are at least as valid as the traditional examination.

2. There is no reason to believe that the newer objective tests are impotent for the measurement of reasoning and thought in contrast with memory for facts.

3. If recall tests are held to be valid (and there is no evidence to the contrary), recognition tests measure roughly the same abilities or functions.

4. When validities are measured against school marks as a criterion, the correlations are lower than where long objective tests are used as the criterion of validity. Such a finding, however, is in line with the expectancy, since school marks are very unreliable and hence will not support high correlations.

5. Instructions against guessing seem to give more valid results than where pupils are directed to guess.

6. When validity coefficients are corrected for attenuation (errors due to unreliability of measurement), the resulting values are high, showing that true-false, multiple-choice, and recall tests measure roughly the same abilities.

[1]Two sets of measures cannot correlate higher than the square root of the product of their reliability coefficients, except as a matter of chance. The values found by Charles are very close to such limits.

COMPARATIVE RELIABILITIES

Toops's investigation. Toops seems to have made the first important contribution on the subject of the relative reliabilities of various types of objective test items.[1] He made first a fifty-item recall test over general information, of which the following are samples:

1. What letter designates the note on the bottom line of the staff in music? *Ans.* E.

50. In what city is the Smithsonian Institute? *Ans.* Washington.

The items were then changed into five-response recognition types, and then to true-false; e. g., Item 1 became:

(*Recognition*) What letter designates the note on the bottom line of the staff in music? *Ans.* (A, E, G, B, C).

(*True-false*) The letter G is the note which is on the bottom line of the staff in music. True False

Table 32 reproduces Toops's results (Table II, p. 49 of Toops's monograph).

TABLE 32

COMPARISON OF RELIABILITY COEFFICIENTS OF THE RECALL, RECOGNITION, AND TRUE-FALSE TESTS

	RECALL	RECOGNITION	TRUE-FALSE
Reliability (r_{11}) of halves, 124 cases. Two forms of 25 each	.448	.385	.340
Reliability of two 50-question sets. (Brown's formula, $n=2$)	.618	.556	.507
Average time in minutes to do 50 questions	6.9	5.6	3.6
Number of questions per unit of recall time	1.00	1.23	1.92
Number of sets of 25 questions to get equal reliability of .618	2.00	2.60	3.14
Reliability of Form A with Form B when 6.9 minutes of examination time are used	.618	.607	.664

[1] H. A. Toops, "Trade Tests in Education," *Teachers College Contributions to Education*, No. 115 (New York: Columbia University, 1921), especially pages 39-62.

The following conclusions seem justified from Toops's data:

1. In order of decreasing reliability, the tests stand in the order: recall, recognition (five-response), and true-false, *when fifty-item tests are compared.*

2. The average working times needed were: recall, 6.9 minutes; recognition, 5.6 minutes, and true-false 3.6 minutes.

3. In the time needed for 100 recall items, 123 recognition items can be answered, and 192 true-false items can be responded to.

4. When reliabilities are estimated for *equal working times* (6.9 min.), the orders of rank are: true-false (.664); recall (.618); and recognition (.607). Equal working times afford, it seems, the best basis for such comparisons.

The work of Toops on college students has lead directly or indirectly to the studies of Ruch and Stoddard, Wood, Paterson, DeGraff and Ruch, Brinkley, Charles, and others. These studies will next receive attention.

The investigation of Ruch and Stoddard.[1] These authors selected a set of 100 information items covering the general field of history and the social sciences, suitable in difficulty for twelfth-grade pupils. These items were next broken by chance into two approximately equal "forms," designated as Form A and Form B. The items were then adapted to each of the following types:

1. Recall
2. Recognition, 5-response
3. Recognition, 3-response
4. Recognition, 2-response
5. True-false

Thus, each of the 100 items appeared in five different type-forms. Two items (Form A only) are given on the next page in all five types.

[1]Quoted with changes and deletions from G. M. Ruch, *The Improvement of the Written Examination* (Chicago: Scott, Foresman and Co., 1924), pp. 107-114. For a fuller report see: G. M. Ruch and G. D. Stoddard, "The Comparative Reliabilities of Five Types of Objective Examinations," *Journal of Educational Psychology*, Vol. XVI (1925), pp. 89-103.

I. Recall, Form A

1. The American Revolution began in the year .....
50. Passports are issued by the Department of .....

II. Five-Response, Form A

1. The American Revolution began in (1) 1762 (2) 1775 (3) 1783 (4) 1789 (5) 1812 .....
50. Passports are issued by the Department of (1) State (2) Commerce (3) Interior (4) War (5) Labor .....

III. Three-Response, Form A

1. The American Revolution began in (1) 1762 (2) 1775 (3) 1789 .....
50. Passports are issued by the Department of (1) State (2) Commerce (3) Interior .....

IV. Two-Response, Form A

1. The American Revolution began in (1) 1762 (2) 1775 .....
50. Passports are issued by the Department of (1) State (2) Commerce .....

V. True-False, Form A

1. The American Revolution began in 1775. *True* *False*
50. Passports are issued by the Department of Commerce. *True* *False*

"In order to keep practice effects at as nearly a minimum as possible, it seemed inadvisable to attempt to have each pupil take the two forms in all five ways. For this reason all pupils were given the recall type, Form A, followed directly by Form B, and then one day later were given the same items in *one other* type-form. To administer the tests in this way, the total group of twelfth-grade pupils was broken into four sub-groups, designated as groups A, B, C, and D by a strictly alphabetical division. The senior classes of about fifteen Iowa high schools were arranged in alphabetical order, keeping the schools separate. The first one-fourth in the alphabet, all schools combined, were called Group A, the second one-fourth, Group B, and so on for Groups C and D.

"It will thus be seen that the groups were random samplings with every school contributing equal numbers to each group. Since more than five hundred pupils were involved, the sub-groups can be accepted as equal in ability for all practical purposes. The sub-groups numbered about one hundred thirty-five pupils. The following tabulation will make these points clearer.

Group	No.	Day 1	Day 2
A	137	Recall A and B	5-Response A and B
B	134	Recall A and B	3-Response A and B
C	135	Recall A and B	2-Response A and B
D	133	Recall A and B	True-false A and B

"The recall type was given first for two reasons: first, because it is least suggestive of the correct answers and hence produces smaller practice effects on later tests; and second, in order that all four groups might take one test in common as a check on the equivalence of abilities of the groups.

"The reliability coefficients are given in Table 33.

TABLE 33

Reliability Coefficients of the Five Types of Examination

(*a*)	(*b*)	(*c*)	(*d*)
Type	Form A *vs.* Form B (i.e., 50 items *vs.* 50 items)	Reliability of 100 Items (by the Spearman-Brown formula)	*N*
Recall	.81 ± .010	.90	562
5-response	.80 ± .021	.89	137
3-response	.60 ± .037	.75	134
2-response	.74 ± .027	.85	135
True-false	.56 ± .040	.71	133

"The figures in column (*c*) were calculated by means of the Spearman-Brown formula by using $n = 2$, as an estimate of the reliability of 100 items.

" . . . The times needed to complete one hundred items were kept to the nearest half-minute, thus making it possible to determine: (1) the relative rapidity of administration of each type, and (2) the reliability per unit of working time. Table 34 gives such calculations.

"The figures in column (*b*) of Table 34 are the average times needed by the groups of 135 pupils to complete (i.e., attempt) 100 items. The numbers in column (*c*) are the ratios of the time of the recall test to the times of each of the recognition tests, thus 18.7/16.0 = 1.17, 18.7/13.5 = 1.39, etc. This means that 117 five-response items can be given in the same length of time needed for 100 recall items, 18.7 minutes; 139 three-response items can be given in the time needed for 100 recall items, etc. The values in column (*d*) are brought forward from column (*c*) of Table 33.

TABLE 34

RELATIVE TIMES NEEDED FOR EACH TYPE OF EXAMINATION AND RELATIVE RELIABILITIES PER UNIT OF WORKING TIME

(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
TYPE	TIME IN MINUTES TO COMPLETE 100 TEST ITEMS	ITEMS THAT CAN BE GIVEN IN 18.7 MINUTES (Recall as base)	RELIABILITY OF 100 ITEMS	RELIABILITY PER 18.7 MINUTES WORKING TIME
Recall.....	18.7	100	.90	.90
5-response.....	16.0	117	.89	.90
3-response.....	13.5	139	.75	.81
2-response.....	11.4	164	.85	.90
True-False.....	10.2	183	.71	.82

"The question at once presents itself whether equal working times would result in equal reliabilities of the several types. If, for example, 183 true-false items can be answered in 18.7 minutes (the time needed for 100 recall items), what are the comparative reliabilities of 183 true-false items and

100 recall items? We have already seen the usefulness of the Spearman-Brown formula for arriving at such comparisons. Taking the coefficients of column (*d*) of Table 34 and using *n* equal to 1.17, 1.39, 1.64, and 1.83, in turn, we obtain the values in column (*e*). These might be read as follows: 'the reliability of type . . . . for 18.7 minutes working time.' It will be seen that the original differences are cut down to such an extent that at least two of the types prove to be as satisfactory as the recall under equal time limits. These are the 5-response and 2-response forms. The true-false and 3-response seem somewhat inferior."

The following conclusions seem to follow from the data gathered by Ruch and Stoddard:

1. For a *constant number of items*, the five tests rank as follows in order of decreasing reliability: recall, five-response, two-response, three-response, and true-false.

2. When tests of *equal working times* are compared, the differences are small but favor the recall, five-response, and two-response.

3. There is no apparent reason why the two-response test proved more reliable than the three-response in this study, the a priori expectancy being to the contrary.

4. There is a reasonably close agreement between this study and that of Toops previously reported.

Wood's investigations. Wood has made numerous investigations of reliabilities of old- and new-type examinations. The following data are selected, as typical, from two of his more recent and extensive studies. Table 35 shows selected reliability coefficients from Wood's findings in these two studies.

Wood's reliability coefficients are somewhat higher than most of those previously reported, for at least two reasons: (*a*) his tests were longer; and (*b*) he used larger numbers of cases and more heterogeneous groups in most cases, especially

TABLE 35

SELECTED RELIABILITY COEFFICIENTS FROM WOOD'S STUDIES OF OBJECTIVE TEST TYPES

SUBJECT	NO. OF ITEMS	NO. OF CASES	ACADEMIC LEVEL	TYPE	RELIABILITY COEFFICIENTS	
					Rights	R—W
French*	50	100	College	T-F	.83†	.80†
Law (Pleading)	90	74	College	T-F	.83†	.77†
Law (Property)	100	100	College	T-F	.75†	.76†
Law (Equity)	70	100	College	T-F	.66†	.65†
French, Part I	100	2000	J.H.S.	5-R	.94**	
French, Part II	60	2000	J.H.S.	5-R	.95**	
French, Part III	60	2000	J.H.S.	Completion	.96**	
French, Parts I to III	220	2000	J.H.S.	5-R and Com.	.97**	
Spanish, Part I	100	2000	J.H.S.	5-R	.93**	
Spanish, Part II	60	2000	J.H.S.	5-R	.92**	
Spanish, Part III	65	2000	J.H.S.	Completion	.94**	
Spanish, Parts I to III	225	2000	J.H.S.	5-R and Com.	.97**	

*B. D. Wood, "Studies in Achievement Tests," *Journal of Educational Psychology*, Vol. XVII (1926), pp. 1-22; especially pp. 6-7.

†Averages of four coefficients.

**B. D. Wood, *New York Experiments with New-Type Modern Foreign Language Tests* (N. Y.: The Macmillan Co., 1927), p. 40. Quoted by permission of the Macmillan Co.

in the French and Spanish examinations reported last in Table 35.[1]

Wood's studies do not show direct comparisons of various types of items based on the *same* subject-matter, except roughly so in the French and Spanish tests administered to

[1]The size of the coefficient of correlation depends, in part, upon the range of the values correlated. Kelley calls this phenomenon "range of talent." Another name is "heterogeneity." (See Chapter XV.) Since the 2000 pupils taking the French and Spanish examinations were junior high-school students, it is reasonable to suppose that they represented a greater range of individual differences than was true of the groups taking the first-mentioned French and the Law examinations, the latter being college examinations. Between junior high-school and college there is a great deal of selective elimination, as is well known. This would tend to make the reliability coefficients higher in the case of the younger pupils.

It should be noted that the Toops study dealt with college groups, the Ruch-Stoddard study with high-school seniors, and the two studies of Wood with junior high-school and college students. The three studies show somewhat different results, but many of these differences are to explained by the nature of the groups tested (range of talent or heterogeneity); although some consideration must be given to the different numbers of cases, the different school subjects concerned, and to the types of tests employed.

2000 New York junior high-school pupils, where 60- or 65-item completion tests may be compared with 60- or 100-item multiple-choice tests (five-response). The differences are slightly in favor of the completion test, but not significantly so. This finding is in substantial agreement with the results of Toops and of Ruch and Stoddard. The fact that tests of from 50 to 100 items yield reliabilities of from .66 to .96 is reassuring.

The investigation of DeGraff and Ruch.[1] Using, with modifications, the procedure of Toops and Ruch-Stoddard, these authors devised 200 simple recall items covering the general field of American history. These 200 items were then broken by chance into two forms of 100 items each, designated as Form A and Form B. Each form was then "translated" into items of the seven-response, five-response, three-response, two-response, and true-false types, respectively (in the same manner as was described for the experiment of Ruch and Stoddard).

A total of 2453 pupils took Recall, Form A, the first day of the experiment, followed by Recall, Form B, the second day. On the third and last day the total group was divided into ten sub-groups, *by chance*, as follows:

1. 7-Response "Do Not Guess"
2. 7-Response "Guess"
3. 5-Response "Do Not Guess"
4. 5-Response "Guess"
5. 3-Response "Do Not Guess"
6. 3-Response "Guess"
7. 2-Response "Do Not Guess"
8. 2-Response "Guess"
9. True-false "Do Not Guess"
10. True-false "Guess"

The "Guess" and "Do-Not-Guess" sub-groups (of each pair) received the same test items. One group was instructed emphatically to guess at all items when in doubt and to leave no items blank. The other group was directed

[1]Reported in full in G. M. Ruch *et al, Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Company, 1926,) pp. 54-88. For a brief account see the *Journal of Educational Psychology*, Vol. XVII (1926), pp. 368-375.

to omit all doubtful items and under no circumstances to guess. Table 36 shows the obtained reliability coefficients.

TABLE 36

RELIABILITY COEFFICIENTS FOR SIX TYPES OF OBJECTIVE TESTS COVERING THE SAME INFORMATION (DEGRAFF AND RUCH)

TEST	"GUESS" INSTRUCTIONS		"DO NOT GUESS" INSTRUCTIONS	
	Rights	Rights minus Wrongs	Rights	Rights minus Wrongs
Recall (.950)	....	....	....	....
7-Response	.800	.839	.886	.907
5-Response	.864	.902	.862	.882
3-Response	.837	.858	.886	.890
2-Response	.745	.864	.859	.843
True-false	.641	.780	.885	.837

The average working times were very different for the different types of tests. Table 37 shows the facts.

TABLE 37

AVERAGE TIME IN MINUTES TO ANSWER 200 ITEMS OF THE VARIOUS TYPES AS FOUND BY DEGRAFF AND RUCH

TYPE	TIME	NO. OF CASES
7-response(g)* (200 items)	45.9	212
7-response(n)† (200 items)	40.8	206
5-response (g) (200 items)	42.2	214
5-response (n) (200 items)	37.6	262
3-response (g) (200 items)	38.2	239
3-response (n) (200 items)	37.6	219
2-response (g) (200 items)	34.6	207
2-response (n) (200 items)	34.5	251
True-false (g) (200 items)	33.0	227
True-false (n) (200 items)	30.5	244
Recall A (100 items) 25.2 Recall B (100 items) 24.8 = 50.0		Cases 2200

*(g) indicates the tests taken under instructions to guess.
†(n) indicates the tests taken under instructions not to guess.

Table 38 shows the reliability coefficients brought to the *common* base of the time required to answer 100 recall items by means of the Spearman-Brown prophecy formula. The original or actual coefficients (those of Table 36) are also shown for the sake of comparison. The method of estimating reliability coefficients upon a basis of equal working times has already been described in connection with the abstract of the study by Ruch and Stoddard.

TABLE 38

RELIABILITY COEFFICIENTS FOR THE DEGRAFF-RUCH INVESTIGATION AS ESTIMATED FOR EQUAL WORKING TIMES BY MEANS OF THE SPEARMAN-BROWN FORMULA (The basis is that of the time needed by the average pupil for answering 100 recall items.)

TYPE OF TEST	UNCORRECTED FOR CHANCE		CORRECTED FOR CHANCE	
	Estimated by Spearman-Brown Formula	Original	Estimated by Spearman-Brown Formula	Original
7-response (g)	.815	.800	.851	.839
7-response (n)	.901	.886	.920	.907
5-response (g)	.884	.864	.917	.902
5-response (n)	.893	.862	.908	.882
3-response (g)	.871	.837	.883	.858
3-response (n)	.913	.886	.916	.890
2-response (g)	.805	.745	.900	.864
2-response (n)	.955	.859	.905	.843
True-false (g)	.729	.641	.842	.780
True-false (n)	.925	.884	.892	.837

Correlation of Recall A vs. Recall B	.950
Coefficient of Reliability (Recall A plus Recall B)	.970

The following conclusions seem to be justified by the results of the DeGraff-Ruch experiments:

1. The recall is the most reliable test of the six types studied when an equal number of items are compared.

2. Instructions against guessing when in doubt raise the reliabilities, especially when the scores are corrected for chance by the formula, $\text{Score} = R - \frac{W}{n-1}$.

3. The true-false type proved least reliable, the recognition types falling in intermediate positions, and the recall best.

4. The average working times were quite unequal, the recall requiring more time than the true-false in about the ratio 50:32, or about 3:2.

5. When tests of *equal working times* are compared through estimates of reliability made possible by the Spearman-Brown formula, none of the recognition types (including the true-false) quite reached the reliability of the recall, although the approach was close enough in most cases to justify the conclusion that *for equal working times recall and recognition types are not greatly dissimilar in reliability* (especially when the instructions are against guessing and scores are corrected for chance).

Charles's Study. The study by Charles[1] gives additional data on five types of objective tests as applied to students in college classes in general psychology. Table 39 shows his results.

TABLE 39

RELIABILITIES FOUND BY CHARLES FOR FIVE TYPES OF OBJECTIVE TESTS IN AN EXAMINATION IN ELEMENTARY PSYCHOLOGY

TYPE OF TEST	NO. OF ITEMS	*N*	SCORE = RIGHTS	SCORE = RIGHTS MINUS WRONGS
Recall	50	747	.60	....
5-Response	50	182	.68	.67
3-Response	50	188	.62	.62
2-Response	50	188	.48	.53
True-false	50	187	.60	.55
Average of the four recognition tests			.595	.592
The probable errors range between .016 and .038				

[1] *Op. cit.*

Charles's reliability coefficients are comparatively low, although it must be remembered that he dealt with short tests (50 items) and with selected college groups. His values are in good harmony with those of Toops, who dealt with roughly similar groups. Charles's results are somewhat out of line with previously reviewed studies in that the recall tests were not noticeably better than the recognition tests (except two-response), and in one case (five-response) the recall tests proved less reliable. Charles's recall tests, on the other hand, were not as highly objective as were those of Toops, Ruch-Stoddard, and DeGraff-Ruch.

Rutledge's investigations. Rutledge[1] made a number of studies on tests given to college classes in psychology. His results are considerably different from those reported elsewhere in this volume, as Rutledge himself points out. Tables 40 and 41 present selected findings from the investigations carried out by Rutledge.

Rutledge comments on his results as follows:

> This correction for length of time allowed for the test does not change the relative order of the reliabilities of the tests. Since these relative reliabilities are so widely divergent from those found by other investigators,[2] an analysis of the number of attempts on each type of question may suggest the causes of the differences.

Using the averages from (our) Table 40, Rutledge calculated the values to be expected by tests of 120 items. (See Table 41.)

Rutledge concludes:

> All of the expected reliabilities are high for thirty minutes of testing and indicate that any one of the three types of question may be used in constructing reliable tests, although the true-false is not as good as multiple choice or completion.

[1] R. E. Rutledge, "The True-False Examination in Elementary Psychology with Suggestions for its Improvement," Ph. D. Thesis, University of California. Unpublished.

[2] Toops found recognition and true-false types of examinations to have equal reliabilities for equal times of testing. Ruch and Stoddard found three-response and true-false tests to be approximately equal in reliability. Brinkley found the order of reliability in American history to be completion, true-false, and multiple choice.

TABLE 40

EXPECTED RELIABILITIES OF 12-MINUTE 40-ITEM TRUE-FALSE, MULTIPLE-CHOICE, AND COMPLETION TESTS (Table VII from Rutledge)

COMPUTED FROM MATERIAL	$r_{\frac{1}{2}\frac{1}{2}}$			EXPECTED r_{12}		
	T-F	Comp.	M.C.	T-F	Comp.	M.C.
1st 40 items....	.49	.70	.79	.66	.82	.88
2d 40 items....	.62	.66	.79	.76	.79	.88
3d 40 items....	.52	.68	.85	.69	.81	.92
Average.....	.54	.68	.81	.70	.81	.89

TABLE 41

EXPECTED RELIABILITIES OF 120 STATEMENTS OF EACH TYPE OF QUESTION (Table IX from Rutlege. Based on average r_{12})

	EXPECTED RELIABILITY
120 True-False Statements.....	.87
120 Completion.....	.94
120 Multiple Choice.....	.96
Average.....	.92

The study of Crawford and Raynaldo.[1] These writers found that out of twenty comparisons fifteen favored the old-type test. It is doubtful whether much significance is to be attached to these results, as indeed the authors point out, since the makers of the true-false tests were little skilled in such work. Such tests were also new to the students. Moreover, the numbers of cases were very small in all but a few comparisons, ranging from nine to seventy-four; in ten classes there were fewer than twenty students, and there were only six classes with more than thirty students.

Reliability of matching tests. Ruch, Murdock, and Maupin prepared 120 paired items whereby 120 *men* were to be matched with the same number of *characterizing*

[1]C. C. Crawford and D. A. Raynaldo, "Some Experimental Comparisons of True-False Tests and Traditional Examinations," *School Review*, Vol. XXXIII (1925), pp. 698-706.

phrases. The 120 items were then broken into Forms A and B of 60 items each. Each of the two forms was next prepared as five *groupings*, i. e., sets of 5, 10, 15, 20, and 30 pairs to be matched.[1]

The purpose of this study was twofold: (*a*) To study the change in reliability in going from pairs of 5, 10, 15, 20, to 30. (*b*) To study the change in difficulty in the same series; the increase in difficulty in larger groupings being principally due to lessened opportunity for chance successes, although at least one other factor also enters in a minor degree.

A parallel experiment, but dealing with the matching of *dates* and events was carried out at the same time. The reliability coefficients, and certain other facts, are given here, the discussion of relative difficulties being reserved for a later section of this chapter (page 316).

The sample from Form A on page 305 will help to make the details somewhat clearer. The sample shows the first thirty items (or half) of Form A in the grouping by fives. In the grouping by tens, the first two sets of five pairs were pooled; the same pooling being carried on successively for the fifteen, twenty, and thirty groups.

TABLE 42

COMPARATIVE RELIABILITIES OF MATCHING TESTS OF VARYING GROUPINGS OF PAIRS

GROUPING	DATES-EVENTS				MEN-CHARACTERIZATIONS			
	GRADE 8		GRADE 12		GRADE 8		GRADE 12	
	r	*N*	*r*	*N*	*r*	*N*	*r*	*N*
5's	.86	129	.90	130	.93	161	.94	148
10's	.74	124	.86	121	.77	164	.93	151
15's	.84	121	.80	124	.89	168	.92	146
20's	.79	127	.88	125	.90	159	.98	145
30's	.76	124	.91	124	.88	160	.95	146

[1]For complete details, see G. M. Ruch *et al*, *Objective Examination Methods in the Social Studies*, pp. 89-104.

Matching Test: Men and Characterizations

Directions: Read each *characterizing phrase* and then find the *man* at the left whom the phrase fits best. Record the *number* of the proper man in the parenthesis in front of each phrase. Notice the first item is already filled in correctly. *Each phrase must be matched with a man in the same section.* Work as fast as you can without making mistakes.

FORM A

Men		Characterizing Phrases
		Section 1
1. Thomas H. Benton	(5)	Author of the Declaration of Independence
2. Thaddeus Stevens	( )	For thirty years a senator from Missouri
3. George B. McClellan	( )	An immigrant who worked for political reform
4. Carl Schurz	( )	Leader of Union Army in Peninsula Campaign
5. Thomas Jefferson	( )	Congressman demanding harsh treatment of South
		Section 2
6. Miles Standish	( )	Discoverer of the New World for Spain
7. De Witt Clinton	( )	Spent a fortune to found a colony in America
8. Charles Sumner	( )	Military man of Plymouth, told of by Longfellow
9. Sir Walter Raleigh	( )	Massachusetts senator denouncing "Crime Against Kansas"
10. Christopher Columbus	( )	Governor of New York—promoted the Erie Canal
		Section 3
11. Vasco de Balboa	( )	Governor of Plymouth Colony and Pilgrim leader
12. David Wilmot	( )	Discovered the South Sea, or Pacific Ocean
13. Woodrow Wilson	( )	Offered a "Proviso" concerning slave territory
14. William Bradford	( )	Portuguese explorer who rounded Africa to India
15. Vasco da Gama	( )	The "Great War President"—League of Nations
		Section 4
16. William T. Sherman	( )	U. S. agent to France during X-Y-Z affair
17. Cyrus W. Fields	( )	Invented cylindrical newspaper printing press
18. John Winthrop	( )	Northern general who marched through Georgia
19. Richard Hoe	( )	Laid the first successful Atlantic cable
20. Charles C. Pinckney	( )	Puritan Governor of Massachusetts Bay Colony
		Section 5
21. James J. Hill	( )	Fifth President—"Era of Good Feeling"
22. William McKinley	( )	Orator who denounced "Writs of Assistance"
23. James Monroe	( )	Railroad "King"—builder of Northwest
24. Henry Clay	( )	Third "Martyr President"—shot by anarchist
25. James Otis	( )	Kentucky statesman famous for compromises
		Section 6
26. John Jacob Astor	( )	A South Carolina advocate of nullification
27. Andrew Jackson	( )	The "Little Giant" debating with Lincoln
28. John C. Calhoun	( )	Founded a fur-trading company in Oregon
29. George H. Meade	( )	Hero of New Orleans, first president from West
30. Stephen A. Douglas	( )	Northern general who won at Gettysburg

Go Right on to Page 2 [Items 31-60 appeared on Page 2]

No time limits were set, but each form occupied less than a class period. Unfortunately the data on working times have been lost.

Matching tests seem to be highly reliable, especially the matching of men with characterizing phrases. The lower reliability of the date-events matchings is largely to be explained by the fact that dates are difficult of memory, and the schools are tending (rightly) to minimize such learning.

It cannot well be said that large numbers of pairs to be matched has any great advantage so far as reliability is concerned. We shall see in a later section that the larger groupings do reduce the average scores and hence probably eliminate much guessing. Against this gain there is to be set the greater amount of time needed and the danger of mistakes of carelessness. It may well be that these factors operate to prevent the larger groupings from showing greater reliability, as is the a priori expectancy.

COMPARATIVE WORKING TIMES

Introduction. It would be very useful for the test-maker to know more or less accurately the number of items of the different types which can be given in a stated period of time. Such data could not be used slavishly, since the time requirements probably vary greatly with such facts as: (*a*) the maturity of the pupils, (*b*) mental abilities of the pupils, (*c*) the difficulty of the test, (*d*) the school subject, (*e*) the nature of the test (reasoning or reproduction of facts), etc.

Five previously cited studies have dealt with this question in some detail, viz., the investigations of Toops, Ruch and Stoddard, Brinkley, DeGraff and Ruch, and Charles. These five studies represent very different experimental conditions: grade ranges from seventh grade to college; numbers from 71 to 2200; subjects as varied as general information, history, and psychology; and factual vs. reasoning tests.

The author has previously published experimental data on the actual working times used by pupils in answering tests of different lengths and of differing types of items.[1] These findings were used as a basis of tentative recommendations for proper time allowances. These have been objected to by certain writers. For this reason it has been thought desirable to bring together in one table the available evidence.

Table 43 presents a concise summary of most of the evidence which is easily accessible. The entries for average time and average score are all experimental findings, although it is not to be supposed that each series of tests contained exactly 100 items. It has been necessary to bring them to a common base (100 items). For example, if a given investigator used a 50-item test, his averages and average times were multiplied by two. This assumes merely (*a*) that the pupils would continue to answer at the same rate, and (*b*) items equally difficult and similar in other respects. These assumptions are not particularly dangerous.

There are enormous differences in most or all of the results for any particular type of test. In general Brinkley's results are out of harmony with the others, taken as a whole. The reason for this is unknown. It can hardly be due to sampling, even if his experimental group was small. The probable explanation may center about two facts: (*a*) the longer lengths of many of the statements in his test items, and (*b*) the use of many thought questions (the ratio of thought to information questions was about 1 to 2). Brinkley's questions were, on the average, at least twice the length of those of Toops, Ruch and Stoddard, DeGraff and Ruch, and Charles. This would require much more time for the sheer reading of the questions.

Before attempting to draw up recommendations concerning the rate at which objective tests may be answered, it

[1] *The Improvement of the Written Examination*, pp. 97; 113-114. *Objective Examination Methods in the Social Studies*, pp. 80-84.

TABLE 43

SUMMARY OF AVERAGE WORKING TIME REQUIRED FOR ANSWERING VARIOUS TYPES OF OBJECTIVE TEST ITEMS

(After Toops, Ruch and Stoddard, Brinkley, DeGraff and Ruch, and Charles)

TYPE OF TEST	INVESTIGATOR	SUBJECT	AVER. TIME (100 Items)	AVER. SCORE (100 Items) ("Rights")	GRADE LEVEL	NO. OF CASES
Recall	Toops	Genl. Inf.	13.8[1]	63.1[1,2]	College	124
Recall	Ruch-Stoddard	History	18.7	23.0[2]	Gr. 12	562
Recall	DeGraff-Ruch	History	25.0	27.6[3]	Gr. 7-12	2200
Recall (Word or Phrase Ans.)	Brinkley[7]	History	59.6[4]	41.0[6]	Gr. 12	76
Recall	Charles[8]	Psychology	35.8	26.9	College	747
Completion	Brinkley[7]	History	70.5[4]	42.5[6]	Gr. 12	81
7-Response	DeGraff-Ruch	History	21.7	48.0[3]	Gr. 7-12	206
5-Response	Toops	Genl. Inf.	11.2	70.2[1,2]	College	124
5-Response	Ruch-Stoddard	History	16.0	50.0[2]	Gr. 12	137
5-Response	DeGraff-Ruch	History	18.8	51.5[3]	Gr. 7-12	262
5-Response	Charles[8]	Psychology	25.5	58.9	College	182
4-Response[5]	Brinkley[7]	History	54.4[4]	57.7[6]	Gr. 12	76
3-Response	Ruch-Stoddard	History	13.5	57.0[2]	Gr. 12	134
3-Response	DeGraff-Ruch	History	18.8	58.2[3]	Gr. 7-12	219
3-Response	Charles[8]	Psychology	21.5	70.7	College	188
2-Response	Ruch-Stoddard	History	11.4	67.6[2]	Gr. 12	135

TABLE 43—CONTINUED

TYPE OF TEST	INVESTIGATOR	SUBJECT	AVER. TIME (100 Items)	AVER. SCORE (100 Items) ("Rights")	GRADE LEVEL	NO. OF CASES
2-Response	DeGraff-Ruch	History	17.2	68.4[3]	Gr. 7-12	251
2-Response	Charles[8]	Psychology	19.6	77.5	College	188
True-false	Toops	Genl. Inf.	7.2	78.0[1,2]	College	124
True-false	Ruch-Stoddard	History	10.2	57.8[2]	Gr. 12	133
True-false	DeGraff-Ruch	History	15.2	58.4[3]	Gr. 7-12	244
True-false	Brinkley[7]	History	31.0[4]	45.6[6]	Gr. 12	71
True-false	Charles[8]	Psychology	18.3	67.4	College	189

[1]Calculated by the author from Table I, p. 47, Toops, "Trade Tests in Education." Toops's test contained but 50 items, his means and average working times being estimated for 100 items as a base.

[2]No instructions were given either for or against guessing. (For full data on Toops's study, see *op. cit. supra.*) (For Ruch-Stoddard study, *see* Ruch, *Improvement of the Written Examination*, pp. 107-114, or *Journal of Educational Psychology*, Vol. XVI, 1925, pp. 89-103.)

[3]Average of results for *guess* and *do-not-guess* instructions. For details see Ruch *et al*, *Objective Examination Methods in the Social Studies*, pp. 54-88, or *Journal of Educational Psychology*, Vol. XVII (1926), pp. 368-375.

[4]Calculated by the author from Table II, p. 66 of Brinkley, "New-Type Examinations in the High School." The means and times were brought to the base of 100 from data given by Brinkley.

[5]These tests were considered 4-response; actually, they used from 3 to 5 responses.

[6]Calculated by the author from Table XIA, p. 98 of Brinkley, *op. cit.*

[7]Brinkley's tests were mixed reasoning and information tests in the approximate ratio of 1:2; the tests of the other authors were principally informational. Brinkley's test items averaged much longer than did those of the other authors, and hence required more reading time (it may be supposed). He used "do-not-guess" instructions for the true-false tests.

[8]Unpublished Ph. D. Thesis, University of Iowa, 1926.

will be better to examine the more detailed findings of Ruch and DeGraff as presented in Tables 44 and 45.[1]

The reader is cautioned to keep in mind that the table of working times for recall tests (Table 44) is based upon 100 items, but the times for the recognition types (Table 45) are based upon 200 items. It is further to be remembered that the items used by DeGraff and Ruch were chiefly factual.

TABLE 44

PERCENTILES OF TIME IN MINUTES FOR RECALL TESTS: 100 ITEMS (BY GRADES)

PERCENTILES	FORM A				FORM B				FORM A	FORM B
	GRADE				GRADE				All Grades	All Grades
	7th	8th	11th	12th	7th	8th	11th	12th		
100	44.5	43.0	44.0	44.0	45.0	47.5	40.5	40.5	44.5	45.5
90	30.2	33.4	34.5	36.1	30.3	34.3	32.9	34.0	34.0	33.2
80	27.4	30.6	31.1	32.7	26.4	30.7	30.1	31.2	30.7	30.2
70	25.5	28.3	29.2	30.5	24.5	28.4	28.1	29.7	28.0	27.9
60	23.6	26.5	27.3	28.4	22.5	26.7	25.5	27.2	26.6	25.7
50	22.0	25.1	25.0	26.7	21.1	24.9	24.4	25.5	24.8	24.2
40	20.5	23.6	23.8	25.5	20.0	22.9	23.0	24.2	23.2	22.5
30	19.3	21.9	22.0	23.0	18.7	21.1	21.6	22.6	21.0	21.0
20	17.5	19.7	20.1	21.3	17.5	19.7	20.0	21.5	19.6	19.0
10	15.3	17.6	18.2	18.7	15.9	17.3	18.2	19.5	17.4	17.3

The following comments are based principally on Tables 44 and 45, although Table 43 has been kept in mind in drawing conclusions.

RECALL TESTS

1. The slowest pupils answered recall items at a rate of more than two items per minute (40-45 minutes for 100 items).

2. Seventh- and eighth-grade pupils are not markedly slower than high-school pupils in answering the tests.

[1] *Objective Examination Methods in the Social Studies*, p. 81.

3. The figures showed that 90 per cent of the pupils finished the 100 recall questions in from thirty to thirty-five minutes, or at a rate of about three per minute.

4. The average pupils (50 percentile) answered the recall items at a rate of about four per minute (twenty-one to twenty-seven minutes for the 100 items). The elementary-

TABLE 45

PERCENTILES OF TIME IN MINUTES FOR RECOGNITION TYPES: 200 ITEMS

PERCENTILES	RESPONSE TYPES								TRUE-FALSE	
	7 (g)	7 (n)	5 (g)	5 (n)	3 (g)	3 (n)	2 (g)	2 (n)	(g)	(n)
100	60.0	60.0	60.5	60.0	60.0	60.5	60.0	60.0	55.0	56.0
90	55.4	50.7	52.6	48.1	48.4	48.0	45.0	43.6	44.2	39.1
80	50.4	48.2	48.6	44.1	45.1	45.2	40.7	40.3	39.0	36.2
70	48.1	45.3	45.7	41.8	42.0	41.9	38.8	38.6	36.9	34.4
60	45.4	42.4	44.6	40.1	40.2	39.7	36.4	36.3	34.5	32.4
50	43.9	40.4	42.0	38.8	38.6	37.4	34.8	35.1	32.8	30.6
40	41.0	38.8	40.2	38.2	36.4	35.3	31.9	32.9	29.8	28.7
30	39.2	36.4	38.6	36.2	34.6	33.1	29.6	30.9	28.8	27.4
20	37.7	33.1	36.1	34.9	32.4	30.6	26.4	28.8	26.2	25.3
10	30.4	29.4	32.4	31.1	29.1	27.4	23.7	25.2	24.6	21.7

school pupils finished somewhat sooner than the high-school pupils, but this is to be explained by the fact that the former did not attempt as many questions.

5. It should be noted that these figures are based upon the times needed to go through the test (answering the questions which the pupils knew) and *not* on the time needed to answer *all* 100 questions. Moreover, these recall tests were very difficult, as is attested by the fact that the average recall score was about twenty-eight correct out of a possible 100. The strictly average pupil probably actually attempted not more than half of the items, and very few scores above 90 were found even on the best papers.

6. If the recall test had been of more nearly "ideal" difficulty (one whose average score is about 50 per cent of the maximum), the pupils would have attempted more

items, but against this fact we can set the opposing one that easier items would be answered more quickly.

7. For two reasons it is not ideal to base recommendations on the times needed by the slowest pupils: (*a*) such times represent wasted time to some degree, as inferior pupils are prone to "putter around" almost indefinitely at such tasks; and (*b*) a slight premium on speed is justifiable. In making standard tests it is often the practice to set the time limits so that 90 per cent can attempt all items within their power. Using the data for nintieth percentiles, we can say that:

(*a*) In junior and senior high-school classes, three recall items per minute is not an excessive requirement provided the items are fairly short and of a factual character. For reasoning tests, one or two items per minute should be a reasonable number.

(*b*) It will be wise to increase these time allowances in the lower grades.

Recognition Tests

1. The slowest pupils answered these at a rate faster than three per minute.

2. Four to five multiple-choice or true-false items were handled per minute of working time, approximately 90 per cent of the pupils having time to attempt all items within their power.

3. The average pupil answered five or six multiple-choice (three to five responses) or true-false items per minute.

4. Some time was saved by instructing pupils to omit items rather than to guess when in doubt.

5. Using the values in the ninetieth percentile row of Table 45, it seems reasonable to conclude (for tests of the type employed by DeGraff and Ruch) that:

(*a*) Four items per minute is a reasonable expectancy for upper-elementary and high-school pupils for multiple-choice tests (three to five responses) and true-false tests.

(*b*) If thought questions are used, two or three items per minute is a safer practice. (This "squares" fairly well with Brinkley's results as well.)

(*c*) Three items per minute would seem to be safe for pupils in grades four to six, although this recommendation is inferred from the working rates of upper-grade pupils.

The foregoing statements are very conservative in the light of the evidence. Brinkley's findings alone might be used as an argument against them. The Toops and Ruch-Stoddard studies indicate much higher rates of work.[1]

A further comparison. It has been stated that because of the very unequal working times needed for various types of objective tests, comparison upon a basis of the number of items which can be completed in a given length of time is fairer than upon a basis of equal numbers of items. It has already been found that 100 recall items can be given in a class period of from forty to sixty minutes. The following data, assembled from several sources and brought to a common base, shows something about the relative working times needed for recall, five-response, and true-false tests.

The agreement is very close indeed when the wide differences in experimental procedures are considered.

TABLE 46

NUMBER OF ITEMS WHICH CAN BE GIVEN IN THE TIME NEEDED FOR 100 RECALL ITEMS

AUTHORITY	RECALL	5-RESPONSE	TRUE-FALSE
Toops	100	123	192
Ruch-Stoddard	100	117	183
DeGraff-Ruch	100	133	161
Brinkley*	100	...	192
Charles	100	140	189

*Brinkley's "Word or Phrase Answer" was considered as simple recall.

[1]The foregoing recommendations are cut almost in half from those originally published in the author's *Improvement of the Written Examination* (p. 97), those being based upon the Toops and Ruch-Stoddard studies which were the only ones in print in 1924. As has been mentioned, objections have been taken to the author's earlier figures. The earlier figures *were* experimental findings. Repetition of the experiments on a large scale by DeGraff and the author has lead to the more conservative recommendations of the present volume.

COMPARATIVE DIFFICULTIES

Introduction. Table 43 on pages 308-9 of this chapter presented certain facts about the comparative difficulties of recall, recognition (multiple-response), and true-false tests. This section presents six studies in some detail.

Toops's results. Toops, as we have seen, gave three versions of the "same" items to college students. The following table is taken from Table I, p. 47, of "Trade Tests in Education," each average being the result when a given type was taken first (Toops gave all three tests in each possible order), in order to eliminate practice effects.

TABLE 47

AVERAGE DIFFICULTIES OF RECALL, 5-RESPONSE, AND TRUE-FALSE TESTS AS FOUND BY TOOPS (The number of items is 50.)

TYPE	AVERAGE SCORE	NO. OF CASES
Recall	29.7	76
5-Response	33.4	26
True-false	35.8	22

Results of Ruch and Stoddard. Table 48 is a comparison of five types of tests these authors studied.[1]

TABLE 48

AVERAGE DIFFICULTIES OF RECALL, MULTIPLE-RESPONSE AND TRUE-FALSE TESTS AS FOUND BY RUCH AND STODDARD

TYPE	NO. ITEMS PER FORM	FORM A	FORM B	NO. OF CASES
Recall	50	12.2	10.8	562
5-Response	50	27.2	22.8	137
3-Response	50	30.6	26.4	134
2-Response	50	35.6	32.0	135
True-false	50	30.1	27.7	133

Brinkley's results. Brinkley ("New-Type Examinations in the High School") obtained the averages shown in Table 49. These results are abstracted from Brinkley's Table

[1] *Improvement of the Written Examination*, p. 112.

XI*a*, p. 98. The averages shown here are for tests of equal working times (thirty-one minutes in each case).

TABLE 49

BRINKLEY'S FINDINGS ON THE COMPARATIVE DIFFICULTIES OF CERTAIN TYPES OF OBJECTIVE TEST ITEMS

TYPE	AVERAGE
True-false	45.6
Multiple-response (3, 4, and 5 responses)	32.9
Completion	18.7
Word or phrase answer (a simple-recall test)	21.3
Arrangement	23.8

It is to be noted that Brinkley's results are not directly comparable with those of Toops, Ruch and Stoddard, Charles (to follow), and DeGraff and Ruch (to follow) because *unequal* numbers of items and *different* items are compared.

Results of DeGraff and Ruch. Table 50 is taken from *Objective Examination Methods in the Social Studies*, p. 79, with changes. This investigation has already been described.

TABLE 50

COMPARATIVE DIFFICULTIES OF OBJECTIVE TEST ITEMS AS REPORTED BY DEGRAFF AND RUCH (AVERAGES)

TYPE OF TEST	RECOG. A (Uncorrected for Chance)	RECOG. A (Corrected for Chance)	RECOG. B (Uncorrected for Chance)	RECOG. B (Corrected for Chance)
7-response (g)*	50.0	41.5	39.6	32.6
7-response (n)†	44.9	40.0	37.2	33.1
5-response (g)	54.2	43.4	45.5	35.4
5-response (n)	48.8	42.3	42.1	36.4
3-response (g)	62.2	43.6	55.5	36.6
3-response (n)	54.1	41.9	48.2	36.1
2-response (g)	71.7	43.6	67.2	37.1
2-response (n)	65.1	45.8	60.3	40.2
True-false (g)	65.8	32.3	61.3	26.0
True-false (n)	51.0	30.8	47.6	26.8

*(g) indicates the tests taken under instructions to guess.
†(n) indicates the tests taken under instructions not to guess.

Charles's study. Table 51 summarizes the results of the unpublished study by Charles already cited.

TABLE 51

RESULTS OBTAINED BY CHARLES ON THE "SAME" ITEMS IN GENERAL PSYCHOLOGY WHEN APPLIED AS FIVE TYPES OF OBJECTIVE ITEMS

TYPE	AVERAGES		NO. OF CASES
	Uncorrected for Chance ("Rights")	Corrected for Chance $\left[R-\frac{W}{(n-1)}\right]$	
Recall	26.9	....	747
5-Response	58.9	48.4	182
3-Response	70.7	56.5	188
2-Response	77.5	55.5	188
True-false	67.4	38.1	189

Ruch, Murdock, and Maupin on matching tests. For the rough outline of the procedure in this investigation, the reader is referred to pages 303-5 of the present chapter. Table 52 shows the average scores for the two matching tests.

TABLE 52

RELATIVE DIFFICULTIES (AVERAGE SCORES) FOR MATCHING TESTS FOR VARYING GROUPINGS (Form A, only; 60 pairs of items)

GROUPING	DATES-EVENTS				MEN-CHARACTERISTICS			
	Grade 8		Grade 12		Grade 8		Grade 12	
	M	*N*	*M*	*N*	*M*	*N*	*M*	*N*
5's	9.7	129	16.8	130	35.0	161	44.7	148
10's	6.4	124	11.5	121	26.0	164	36.7	151
15's	6.1	121	10.4	124	22.1	168	32.3	146
20's	5.2	127	9.5	125	20.0	159	29.9	145
30's	5.0	124	9.9	124	16.3	160	26.2	146

There is a steady decrease in the average scores of both tests as we move in the direction of the large groupings, thus

showing a decrease in the amount of the score due to chance. The dates-events test was far too difficult, the men-characteristics test being of about the proper difficulty.

On the whole the evidence is ambiguous, but logical considerations point toward the use of from ten to twenty pairs as being a fair compromise among the several factors involved.

CHAPTER XII

CHANCE AND GUESSING IN RECOGNITION TESTS

Two approaches to the problem of chance effects in test scores. That most forms of recognition tests are open to the effects of chance and guessing is evident. The extent to which such effects are dangerous is as yet unsettled. There is a considerable literature on chance effects in test scores, especially in the case of the true-false test. These discussions fall into general groups:

(*a*) A priori or mathematical considerations of chance and guessing from the standpoint of the mathematical theory of probability.

(*b*) Experimental studies of the comparative reliabilities, validities, difficulties, etc., of tests subject to and not subject to guessing.

There is no intention here of decrying the merits of the published works of the first mentioned type, although it must be admitted that the author is biased in favor of accepting the results of actual experimentation whenever a priori and experimental results seem to be opposed. In spite of what has been said, it is nevertheless true that many writers on the chance element in true-false tests have misunderstood the implications of the theory of probability in discussing the subject.

Certain writers have chosen to regard the situation of a pupil in taking a true-false test as being analogous to such *chance* situations as drawing by lottery from a container holding black and white buttons (or other objects) or of tossing coins. Under certain circumstances the answering of true-false tests may be a matter of pure chance, but often

it is not. We should draw a distinction at this time between *guessing* and *pure chance* (in the sense that a pupil has no better basis of choice than something equivalent to tossing a coin or drawing black and white balls from a box).

THE MATHEMATICS OF CHANCE APPLIED TO TESTS

Pure chance and guessing contrasted. As a pupil goes through a true-false (or other multiple-choice) test and faces a decision on each successive item, in turn, we can recognize his responses as falling into several roughly distinguishable categories:

(*a*) Items on which he is absolutely sure of the correct response.

(*b*) Items on which he is not entirely certain, but which he answers without serious doubt of the correctness of his responses.

(*c*) Items on which he is in grave doubt. He has a "feeling" (or often, merely a "hunch") that a certain response is correct.

(*d*) Items on which he is totally ignorant, and on which *any* response, so far as he can tell, is a matter of pure chance. In such a case the alternatives are (1) to guess or (2) to omit.

To these we must add a fifth category:

(*e*) Items which are answered in good faith, but are wrong, not due to chance, but because the pupil is *actually misinformed*. These are not guessed responses in any legitimate meaning of the word.

It is logical to assume that only items of the (*d*) category obey the mathematical theorems of probability in their distributions of correct and incorrect responses.

Weidemann, as previously quoted, has commented upon the *specific determiner* as potent in "casting the die" at times when the pupil is in real doubt. When some clue is afforded by the wording of the item, that item falls into category (*c*) rather than (*d*) as listed above.

It is permissible, perhaps, to anticipate later discussions by pointing out that the instructions may be phrased so as to encourage pupils to omit completely items of type (*d*) rather than to guess, and hence much pure guessing may be obviated.

Misunderstandings of the implications of the mathematical theory of probability. One writer[1] has criticized the two-response test in no uncertain terms as follows:

> If one holds that by subtracting the wrong from the right answers one eliminates the guessing factor, no more and no less, one must assume that individuals in such tests *always guess an even number of times;* for if they should guess an odd number of times, the effect of guessing could not be wholly eliminated by this method. One must further assume that if an individual happens to guess wrong the first time, *his second guess must be right, his third wrong, his fourth right,* and so, guessing right and wrong alternately; for if he should guess wrong or right twice in succession, the method could not eliminate the guessing effect. That the law of chance does not operate even approximately this way can easily be demonstrated. (p. 236.)
>
> But if one argues that the guessing factor and no more is eliminated by subtracting the wrong answers from the right, one must still further assume that *every wrong answer is a guess;* otherwise, if the wrong answers are subtracted, they will cancel not guesses but actual achievements. That only guesses cause wrong answers no one would care to assume." (p. 238.)

This writer's final conclusion is:

> What then does the final score of such tests represent? No one knows. That it cannot even approximately represent real ability or actual achievement has been shown. Nothing important, therefore, should be done on the basis of the score. (p. 239.)

(The reference here is to a table showing the results of drawing from a bag containing thirty-five white and thirty-five black buttons.)

With Hahn's second paragraph the author has no particular quarrel if Hahn had in mind the responses labelled

[1] H. H. Hahn. "A Criticism of Tests Requiring Alternative Response," *Journal of Educational Research,* Vol. VI (1922), pp. 236-240.

(*e*) in our preceding discussion. However, *only a portion* of wrong answers fall into category (*e*); many are undoubtedly to be classified under category (*d*).

McCall, in an early paper,[1] comments on the justice of scoring of true-false tests as follows:

> Let us consider first the reason for expressing a pupil's score as the number correct minus the number wrong. Imagine a pupil who is absolutely innocent of any knowledge of the physical features of the United States. Were such a pupil to take the above test and were he to mark every statement, he would according to the theory of chance mark ten statements correctly and ten incorrectly. The chances of his guessing right or wrong are fifty-fifty or one to one. His score on the above would be:
>
> $$\text{Score} = 10 - 10 = 0.$$
>
> In short, the pupil's knowledge is zero, and the method of computing his score gives him zero. Suppose instead that he knows ten statements and guesses at the other ten. Of the ten guessed at, he would, according to chance, get five correct and five wrong. That is, even though his real knowledge is ten, he will show fifteen correct (10+5) and five incorrect. The method of computing his score brings out his real knowledge.
>
> $$\text{Score} = 15 - 5 = 10.$$
>
> A pupil who marks every statement correctly makes a perfect score, viz.,
>
> $$\text{Score} = 20 - 0 = 20.$$

Like Hahn, McCall does not seem to distinguish between the most probable score and the one which is actually obtained. The mathematics of probability does not assume that *every* individual will be justly scored by the $R - W$ formula, but merely that such a correction yields the *most probable* score, and that the *average* of a large number of individual scores or the score of one individual on a very large number of items should fall measurably close to the value given by the $R - W$ method. Here again, too much faith is attached to the exactness with which probability works. A little experimenting with coin tossing is all that is needed to prove the justice of this criticism.

[1] W. A. McCall "A New Kind of School Examination," *Journal of Educational Research*, Vol. V (1920), pp. 33-46.

The replies of Barthelmess and Odell to Hahn's arguments. There have been several refutations of Hahn's arguments, notably those of Barthelmess[1] and Odell.[2]

Combining points raised by Barthelmess and Odell with certain observations of the author, there are four assumptions made by Hahn, viz.:

1. That an individual always guesses an even number of times.
2. That if an individual happens to guess wrong the first time, his second guess must be right, the third wrong, the fourth right, etc.
3. That every wrong answer is assumed to be a guess.
4. That for every wrong answer (due to guessing) exactly the same number were right by guessing.

The first two assumptions reduce to a misunderstanding of the laws of probability. The theorems summarizing chance phenomena imply nothing of the sort. The laws of probability imply merely that *for large numbers of chances* pure guessing will yield right and wrong responses in approximately equal numbers. No order of alternation of such right and wrong guesses is implied, and no question of odd or even numbers of times enters.

Assumption three has some bearing. Barthelmess says, and rightly in the author's opinion (*loc. cit.*, p. 358): "It is true that with the method of scoring used ($R-W$), one assumes that every wrong answer is a guess. It will be guessing unless (*a*) the pupil has learned the fact erroneously, or (*b*) the wording of the question is suggestive."

Barthelmess advances no proof of this statement, but the findings of Weidemann emphasize the effects of certain wordings, and Foster and Ruch, and others, have shown pretty conclusively that some wrong answers are not

[1]H. M. Barthelmess "Reply to a Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. VI (1922), pp. 357-359.
[2]C. W. Odell, "Another Criticism of Tests Requiring Alternative Responses," *ibid.*, Vol. VII (1923), pp. 326-330.

guesses. To the extent that genuine misinformation causes wrong responses, the $R-W$ formula penalizes or over-corrects for chance. Odell goes even further and appears to defend over-penalizing wrong responses (*loc. cit.* pp. 327-328):

> It is a rather generally accepted maxim among educators that when we know something we should know that we know it, and, furthermore, that it is better to know that we do not know something than to think we know it when we do not. Therefore, if the student in question thinks that he knows the correct answers to all the exercises but really gives incorrect answers to five of them (25 in all), a deduction should be made from his positive score of twenty. In other words, according to this system of scoring, a student who knows twenty and does not attempt the other five will receive a higher score than a student who thinks he knows all of them but is mistaken in some cases. This is as it should be.

The reader may not agree with Odell in entirety, but it must be admitted that he has raised an issue that must be settled. We should not, however, confuse Odell's position with that of Hahn who had a different situation in mind.

Barthelmess comes more directly to the point when he says (*loc. cit.*, pp. 357-358):

> In the first place, the best test of any test is correlation with a criterion. If this correlation is satisfactory, we can forget all minor criticisms concerning chance.....
>
> In the second place, no one has a right to insist on perfect reliability. The question at issue is, "Does this method secure more accurate results than other available methods?"

Barthelmess's position will receive considerable support from data to be presented later.

Turning now to Hahn's fourth assumption, viz., "That *just because a certain number of answers are [sic!] wrong, the same number of right answers must be guesses*" (*loc. cit.*, p. 238), we have again a misunderstanding of the laws of probability. The implication of the $R-W$ formula is not that right and wrong guessed responses are exactly equal in numbers but that equality of numbers is the most probable outcome. There is really a considerable difference between

actual values and *most probable* values. Thus, if ten pennies are tossed into the air and allowed to fall at will, the *actual* number of heads (or tails) cannot be controlled or foretold, although the most probable eventuality is five heads. The situation would, theoretically, be the same in a true-false test, *if purely chance answering could be assumed.*

It is quite evident that the $R - W$ formula would not hold *exactly* for every individual pupil if true and false responses follow the same laws as tosses of pennies. The seriousness of this inadequacy of the correction formula will be discussed again as new experimental evidence is brought forward.

Other criticisms of alternate-response tests. West,[1] Asker,[2] Kohs,[3] Kohs and Richards,[4] Richards,[5] Holzinger,[6] and many others have commented at some length on the scoring of two-response tests. Only the briefest mention of these various points of view is possible here.

West gave a nonsense test to be marked "S" or "D" indiscriminately. These test items then were scored by an answer key belonging to another test. The scores thus obtained were corrected by the $R - W$ method. The corrected scores on fifty items ranged from -18 to $+20$, the average being about 1.03, and the resulting distribution almost normal (that representing pure probability). The middle half of the corrected scores fell between -4.2 and $+6.3$, a range of about ten points. West then carried out a somewhat similar experiment with a fifty-item synonym-antonym test, made sufficiently difficult that a great many wrong

[1] P. V. West, "A Critical Study of the Right Minus Wrong Method," *Journal of Educational Research*, Vol. VIII (1923), pp. 1-9.
[2] W. Asker, "The Reliability of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. IX (1924), pp. 234-240.
[3] S. C. Kohs, "High Test Scores Attained by Sub-Average Minds," *Psychological Bulletin*, Vol. XVII (1920), pp. 1-5.
[4] O. W. Richards and S. C. Kohs, "High Test Scores Attained by Sub-Average Minds," *Journal of Educational Psychology*, Vol. XVI (1925), pp. 8-17.
[5] O. W. Richards, "High Test Scores Attained by Sub-Average Minds, III," *Journal of Experimental Psychology*, Vol. VII (1924), pp. 148-156.
[6] K. Holzinger, "On Scoring Multiple Response Tests," *Journal of Educational Psychology*, Vol. XV (1924), pp. 445-447.

guesses (answers) resulted. West's analysis of his results, among other things, pointed to two conclusions: (*a*) more items were guessed right than were guessed wrong, and (*b*) women guessed right somewhat oftener than the men. West concluded that this $R-W$ method ". . . . . is of very doubtful reliability for group testing and especially so for the analysis of individual ability" (p. 8).

Asker reports two experiments somewhat like those of West. He used decks of cards to simulate the situations of two- and three-response tests. Using twenty individuals, he obtained a range of scores from −8 to 10 with an average at zero. When these scores were expressed in per cents, and 70 or 75 was taken as the passing mark, "not one individual was able to pass the test by guessing." Asker next attacks the problem through mathematical analysis by expanding the binomial $(P+Q)^n$, where P is the expectancy of heads by chance and Q is the expectancy of tails by chance. He shows (Table I, p. 236) that the probabilities of obtaining the following scores are as shown in Table 53.

TABLE 53

ASKER'S TABLE OF PROBABILITIES FOR TWO-RESPONSE TESTS (20 ITEMS)

RIGHT	WRONG	CORRECTED SCORES		PROBABILITY: ONE IN
		$R-W$	Per Cents	
20	0	20	100	1,048,576
19	1	18	90	52,429
18	2	16	80	5,519
17	3	14	70	920
16	4	12	60	216
15	5	10	50	68
14	6	8	40	27
13	7	6	30	13
12	8	4	20	8
11	9	2	10	6
10	10	0	0	5.7

For three-response tests corrected by the formula, $R-\frac{1}{2}W$, Asker gives the expectancies (Table II, p. 237) shown below.

TABLE 54

ASKER'S TABLE OF PROBABILITIES FOR THREE-RESPONSE TESTS (20 ITEMS)

RIGHT	WRONG	CORRECTED SCORES		PROBABILITY: ONE IN
		$R-\frac{1}{2}W$	Per Cents	
20	0	20	100	3,486,784,401
19	1	$18\frac{1}{2}$	$92\frac{1}{2}$	87,169,610
18	2	17	85	4,587,874
17	3	$15\frac{1}{2}$	$77\frac{1}{2}$	282,323
16	4	14	70	44,978
15	5	$12\frac{1}{2}$	$62\frac{1}{2}$	7,028
14	6	11	55	1,406
13	7	$9\frac{1}{2}$	$47\frac{1}{2}$	351
12	8	8	40	108
11	9	$6\frac{1}{2}$	$32\frac{1}{2}$	40
10	10	5	25	18
9	11	$3\frac{1}{2}$	$17\frac{1}{2}$	10
8	12	2	10	6.7
7	13	$\frac{1}{2}$	$2\frac{1}{2}$	5.5
6	14	−1	−5	5.5

Asker realized that guessing at *all* items in a test is not the usual situation, and he went on to set up further tables which assumed that certain proportions of the items were known certainly and the remainder only guessed at. If we examine Tables 53 and 54 with care, and make the further assumption that, in a true-false test of 100 or more items perhaps not more than one-third (if that many) are *pure* guesses, the possibility of obtaining any considerable fraction of the total score by chance is seen to be not very great. From Table 53 it is evident that if a pupil guesses at twenty items in a 100-item true-false test, there is but one chance in 216 that he will earn more than ten points more than he deserves.

In general, with true-false tests of sufficient length for adequate reliability, and provided further that the pupil guesses at no more than twenty per cent of the items, there

is little likelihood of errors greater than ten per cent of the total (corrected) score arising often enough to be a serious practical matter. Twenty per cent of guesses may seem a small fraction, but this number (or less) may be obtained by the combination of (*a*) instructions to omit items when in doubt, and (*b*) tests of not too great difficulty. This assertion also takes into account the fact that many so-called "guesses" are not pure guesses but are responses made with a sufficient "fringe of knowledge" to guarantee a high excess of rights over wrongs. Only the pure guesses are to be reckoned with in estimating the probabilities by mathematical formulas.

Kohs and Richards, in the three articles cited, set up even more extensive tables of probabilities, expanding their tables to include four-response tests as well. Space will not permit the reproduction of their numerous tables, although it is only fair to point out that Kohs and Richards are somewhat more skeptical of the adequacy of the chance correction formula than is Asker.

Holzinger (*op. cit.*) gives a simple algebraic proof that the *rights* and *rights minus wrongs* correlate to unity (1.00) *when there are no items omitted.* This is obviously true, although Holzinger seems to be the first to give a mathematical proof. The practical significance of this proof is not altogether clear, as evidence to be presented later shows that guessing at all items may not be a defensible practice.

The foregoing brief abstracts and comments are hardly adequate to the mathematics of the situation, but the author is swayed to omit further reference to the *a priori* and *mathematical* approaches to the problem in favor of actual experimental results obtained under normal classroom conditions. That mathematical and experimental analyses of the chance situation may not agree entirely is to be expected in the light of such considerations as:

(1) Many so-called "guessed" responses represent answers made upon a basis of marginal information or "fringes of knowledge," not very definite or certain, but nevertheless

sufficient to cause an excess of right over wrong responses, and hence do not follow the mathematical expectancies to be derived from the binomial theorem or other basic laws of probability.

(2) Wrong answers, definitely matters of misinformation, occur, and these cannot reasonably be held to follow laws of chance.

(3) There is sufficient evidence (presented later) to prove that much pure guessing can be eliminated by suitable instructions. This fact tends to lessen the errors presumably unavoidable from purely mathematical reasoning.

The binomial theorem applied to chance responses in tests. The serious student of examinations will naturally wish to know the mathematical basis by which Asker arrived at the results given in Tables 53 and 54.

If we may assume that the answering of true-false and other two-response tests reduces to a pure chance situation when the pupil is absolutely ignorant of the correct answers to certain questions, the theoretical expectancy of $R-W$ scores of any given magnitude may be derived from the expansion of the binomial, $(t+f)^n$, where t represents true responses and f stands for false responses.[1]

Any text on college algebra or many elementary treatments of statistical methods will explain the application of the binomial theorem to purely chance situations.[2]

Using t for true responses, f for false responses, and n for the number of items guessed at, the formula for the binomial expansion is given on the next page.

---

[1]Later discussions will show that there is some evidence to the effect that true-false tests behave somewhat differently from two-response recognition tests. In particular, true-false tests seem to be markedly more difficult than two-response tests based upon the same subject-matter, as nearly as the two types of tests can be made to cover the *same* ground. The reason for the greater difficulty of the true-false test may lie in the fact that there is a certain negative suggestion effect in such tests. Moreover, it will be demonstrated later that there are some reasons to think that a true-false test is not a typical two-response test. (See pages 343 to 345.)

[2]H. L. Rietz and A. R. Crathorne, *College Algebra* (New York: Henry Holt and Company, 1919), pp. 93-95. Or,

L. L. Thurstone, *The Fundamentals of Statistics* (New York: The Macmillan Company, 1923), pp. 132-141.

The symbol, !, means the factorial of the number, i.e., the product of all numbers from 1 to that number. Thus, factorial 6 (written 6!) is $1\times2\times3\times4\times5\times6$, or 720.

$$(t+f)^n = t^n + nt^{n-1}f + \frac{n(n-1)}{2!}t^{n-2}f^2 + \ldots$$

$$+\frac{n(n-1)\ldots(n-r+2)}{(r-1)!}t^{n-r+1}f^{r-1} + \ldots + f^n. \quad (\textit{Formula 1})$$

The rth term is:

$$\frac{n(n-1)\,(n-2)\ldots(n-r+2)}{(r-1)!}t^{n-r+1}f^{r-1} \quad (\textit{Formula 2})$$

which may also be written:

$$\frac{n!}{(r-1)!\,(n-r+1)!}t^{n-r+1}f^{r-1} \quad (\textit{Formula 2a})$$

It is actually simpler to find the coefficient of any term (the rth term) by multiplying the coefficient of the preceding term by the exponent of t in that term, and dividing this product by a number one larger than the exponent of f in that term. This procedure is simpler only if each term is being found in order. If a given term is to be found in isolation (the preceding terms not being found) Formula 2*a* above is more convenient.

The first three terms of $(t+f)^{20}$ may be found by Formula 2*a* as follows:

The first term ($r=1$) is:

$$\frac{20!}{(1-1)!\,(20-1+1)!}t^{20-1+1}f^{1-1} = \frac{20!}{20!}t^{20} = t^{20}$$

The second term ($r=2$) is:

$$\frac{20!}{(2-1)!(20-2+1)!}t^{20-2+1}f^{2-1} = \frac{20!}{1!19!}t^{19}f^1 = 20t^{19}f^1$$

The third term ($r=3$) is:

$$\frac{20!}{(3-1)!(20-3+1)!}t^{20-3+1}f^{3-1} = \frac{20!}{2!18!} = 190t^{18}f^2$$

If we continue in the following manner until the last or $(n+1)$th term is reached, we obtain the following table (Table 55) of coefficients of t and the probabilities shown. It is to be noted that there are $n+1$ terms in any expansion of the binomial to the exponent n.

TABLE 55

TABLE SHOWING THE COEFFICIENTS OF EACH TERM AND THE PROBABILITIES OF OCCURRENCE OF T (TRUE) RESPONSES ACCORDING TO THE EXPANSION OF THE BINOMIAL $(t+f)^{20}$

(*a*)	(*b*)	(*c*)
TERM	COEFFICIENTS	PROBABILITY: 1 IN:
1	1	1,048,576.0
2	20	52,428.8
3	190	5,518.8
4	1,140	919.8
5	4,845	216.4
6	15,504	67.6
7	38,760	27.0
8	77,520	13.5
9	125,970	8.3
10	167,960	6.2
11	184,756	5.7
12	167,960	6.2
13	125,970	8.3
14	77,520	13.5
15	38,760	27.0
16	15,504	67.6
17	4,845	216.4
18	1,140	919.8
19	190	5,518.8
20	20	52,428.8
21	1	1,048,576.0
SUM	1,048,576	

Table 55 shows the sum of the 21 coefficients to be 1,048,576. Column (*c*) shows this sum divided by each coefficient in turn in order to express the probability of occurrence.

In order to apply this table to the actual situation of a true-false test, it will help to examine the first and last terms

in the expansion of $(t+f)^{20}$. When expanded the expression reads as follows:

$$t^{20}+20t^{19}f+190t^{18}f^{2}+1140t^{17}f^{3}+4845t^{16}f^{4}+\ldots+190t^{2}f^{18}+20tf^{19}+f^{20}.$$

The first term (t^{20}) represents the situation: 20 true and 0 false. The second term ($20t^{19}f^{1}$) represents the situation: 19 true and 1 false. The third term ($190t^{18}f^{2}$) represents the situation: 18 true and 2 false. The exponents of t in each term represent the numbers of true responses and the exponents of f in each term show the number of false responses. The ratio of the coefficient of any term to the sum of the coefficients of all $(n+1)$ terms is the probability of occurrence of the situation represented by that term.

EXPERIMENTAL INVESTIGATIONS

General considerations. It has been pointed out that there are two general approaches to the problem of chance effects in test scores, viz., (*a*) from the standpoint of the mathematical theory of probability, and (*b*) by actual experimentation. Typical studies of the first type have been considered in the preceding section of this chapter.

Some of the difficulties in the way of applying probability theorems to the chance situation have been mentioned. In the first place, much so-called guessing is not pure chance. The pupil, on the contrary, responds with subliminal or marginal knowledge sufficient to cause him to "guess" right far oftener than wrong. Again, there is the possibility that true-false tests are not situations affording a 50:50 chance for success since there may be positive or negative suggestion effects in such tests. This evidence will be presented in Chapter XIII. In the third place, as we shall soon see, there is some reason to hold that the chance effects are slightly different in true-false and two-response tests proper. Lastly, there is some evidence that guessing should

be discouraged in framing test instructions, and that directions against guessing really help to control the chance factor.

It has been assumed that guessing in recognition tests may be corrected, at least sufficiently well for practical purposes, by the formula: $Score = No.\ Right - \frac{No.\ Wrong}{(n-1)}$, where n is the number of responses presented, and from which one correct answer is to be selected.

For true-false and two-response tests, this formula is:

Score = No. right minus the number wrong, or $S = R - W$

For three-response tests, the formula is:

Score = No. right minus $\frac{1}{2}$ the number wrong, or $S = R - \frac{W}{2}$

Although this general formula appears to be logical for purely chance situations, it is possible to study the actual value of the formula experimentally. There is also the pertinent question of ascertaining whether pupils should be told to guess or not to guess when in real doubt. This also is open to experimental study.

Four questions will serve to state the general issues:

1. Does the $R - \frac{W}{(n-1)}$ formula increase the reliability of test scores when compared with scoring simply the number right?

2. Does the $R - \frac{W}{(n-1)}$ formula increase the validity of test scores when compared with scoring simply the number right?

3. Should pupils be instructed in favor of or against guessing when the answer is entirely unknown (after careful thought)?

4. Does the $R - \frac{W}{(n-1)}$ formula over-, under-, or properly-correct for pure chance successes?

These four questions will be discussed in the following pages. The order of treatment, will be historical rather than topical.

The investigation of the scoring of Army Alpha. The first experimental study of this question seems to be some rough preliminary work done during the preparation of Army Alpha for use in the World War.[1]

The ten tests tried out for the final forms of Army Alpha (of which eight were finally retained) were scored in two ways: (1) number right and (2) number right minus number wrong, for seventy soldiers at Camp Lee. Table 56 shows the results.

TABLE 56

CORRELATION OF RIGHTS AND RIGHTS MINUS WRONGS FOR TEN PRELIMINARY TESTS OF ARMY ALPHA ON 70 CASES. ALL CORRELATIONS ARE AGAINST TOTAL SCORES

TEST NO.	TYPE OF TEST	RIGHTS	$R-W$
1	Oral directions (little chance)	.76	.61
2	Memory for digits (discarded)	.62	.36
3	Mixed-up sentences (true-false)	.66	.72
4	Arithmetic problems (little chance)	.84	.70
5	General information (4-response)	.90	.81
6	Synonym-antonym (same-opposite)	.76	.82
7	Practical judgment (3-response)	.81	.79
8	Number series completion (little chance)	.74	.69
9	Analogies (4-response)	.79	.70
10	Number comparison (later discarded)	.70	.64

It should be noted that the $R-W$ formula applies to two-response tests. The only two-response tests among the ten of Table 56 are numbers 3 and 6. In both of these cases the $R-W$ method gave .06 higher correlation coefficients. In all other cases (where the $R-W$ formula obviously does not apply), the correlation was lowered by the use of corrected scores.

[1] *See* R. M. Yerkes, (Editor), *Psychological Examining in the U. S. Army*, Memoirs of the National Academy of Science, Vol. 15 (1921), pp. 305 and 339.

The effect of corrections for chance on reliability. The next study of the problem seems to be that of Ruch,[1] who studied the effect of corrections for chance by the formula $R-\frac{W}{(n-1)}$ on seven tests of the Terman Group Test of Mental Ability. Table 57 shows the results. The reliabilities were almost uniformly, but slightly, lower when the scores were corrected for chance. This study was based upon an entirely inadequate number of cases (43).

TABLE 57

RELIABILITIES OF SEPARATE TESTS OF THE TERMAN GROUP TEST OF MENTAL ABILITY WHEN CORRECTED AND UNCORRECTED FOR CHANCE

TEST	UNCORRECTED	CORRECTED	TYPE
1. Information	.49 ± .078	.45 ± .082	4-response
2. Best answer	.40 ± .086	.38 ± .088	3-response
3. Word meaning	.67 ± .056	.56 ± .071	2-response
6. Sentence meaning	.53 ± .073	.47 ± .080	Yes-No
7. Analogies	.53 ± .074	.55 ± .072	4-response
8. Mixed sentences	.68 ± .056	.56 ± .071	True-False
9. Classification	.35 ± .091	.41 ± .086	5-response
Average	.52	.48	

Ruch and Stoddard[2] next reported similar results for several types of objective tests over the same subject-matter. Table 58 summarizes the results.

Again, there is little evidence of increased reliability by the use of corrections for chance. (The numbers of cases varied from 133 to 137.) The three-response test was helped by the correction.

Paterson and Langlie[3] next published results very much in harmony with the foregoing. For 111 students (classes of two successive years) taking a true-false test in general

[1] G. M. Ruch, *The Improvement of the Written Examination*, (Chicago: Scott Foresman and Co., 1924), p. 119.

[2] *Op. cit.*, pp. 117-118. The full account of this study appears in: G. M. Ruch and G. D. Stoddard, "Comparative Reliabilities of Five Types of Objective Examinations," *Journal of Educational Psychology*, Vol. XVI (1925), pp. 89-103.

[3] D. G. Paterson and T. A Langlie, "Empirical Data on the Scoring of True-False Tests," *Journal of Applied Psychology*, Vol. IX (1925), pp. 339-348.

psychology, the reliability for "rights" was $0.63 \pm .037$ and for $R - W$ was $0.54 \pm .045$.

TABLE 58

RELIABILITY COEFFICIENTS CORRECTED AND UNCORRECTED FOR CHANCE ELEMENTS

TYPE	UNCORRECTED	CORRECTED $R - \frac{W}{(n-1)}$
5-response	$.80 \pm .021$	$.77 \pm .023$
3-response	$.60 \pm .037$	$.67 \pm .031$
2-response	$.74 \pm .027$	$.68 \pm .031$
True-false	$.56 \pm .040$	$.41 \pm .049$

The evidence surveyed thus far is rather unfavorable to the use of the corrections formula, but it should be pointed out that relatively small populations were used in all these studies. For this, and other reasons, it will be unwise to draw conclusions at this time.

The effect of corrections for chance on validity. During the year 1924-1925 Wood and Ruch, working independently under grants from the New York Commonwealth Fund, attacked the matter of correction for chance from the standpoint of the relative validities of R and $R - W$ scorings. These results were not published until 1926.[1]

Certain facts should be noted about the conduct of these investigations. (Chapter XI gave a short description of the procedures.) Wood used college examinations in law and anatomy courses. His criteria of validity differed somewhat from one course to another but in general included instructors' judgments, essay-examination marks, and other objective tests. The tests used were true-false examinations

[1]B. D. Wood, "Studies of Achievement Tests," *Journal of Educational Psychology*, Vol. XVII (1926), pp. 1-22, 125-139, and 263-269.

G. M. Ruch *et al*, *Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Co., 1926).

G. M. Ruch and M. H. DeGraff, "Corrections for Chance and 'Guess' vs. 'Do Not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology*, Vol. XVII (1926), pp. 368-375.

of from 100 to 200 items. Wood's students were told to omit items when the answering would be a pure guess. DeGraff and Ruch attacked the same problem on the elementary- and high-school level, using a criterion of scores on a simple and highly objective recall test. Against this criterion several multiple-response and true-false tests were tried out. DeGraff and Ruch went one step further than Wood in that the former administered their tests half with instructions to guess, and half with instructions not to guess but to omit when in serious doubt.

Since the section on comparative validities (Chapter XI) summarized both studies, the results are not repeated here. Wood found an average validity coefficient of 0.721 for R scores and 0.769 for $R-W$ scores, a difference of about 0.05. Wood used populations of 100 except in one case where N was 74. Correction also increased the validity of three law examinations by an average of 0.055, the criterion being six essay examinations.

Wood's data on the reliabilities of R and $R-W$ scorings for these same law examinations and also one in French were given in Table 35 of Chapter XI. In general, the former gave slightly more reliable results, and Wood says, "In no case does the score suffer by comparison with the $R-W$ score, and in only one case does $R-W$ compare at all favorably as to reliability."[1]

As before, correcting for chance lowered the reliability slightly, but, more important, *the validity was materially increased.* This study and that of DeGraff and Ruch place a different light on the whole problem of correction.

The results of DeGraff and Ruch were given in some detail in Tables 27 and 36 of Chapter XI. In order to bring both reliability and validity coefficients together in one place, and also in order to see the effects of instructions about guessing, Table 59 is given.

[1] *Loc. cit.* pp. 8-9.

TABLE 59

INTERCORRELATIONS, CORRECTED AND UNCORRECTED FOR CHANCE, FOR ALL TEN TESTS USED

ROW	TYPE OF TEST	VALIDITY RECALL A VS. RECOGNITION A		VALIDITY RECALL B VS. RECOGNITION B		RELIABILITY RECOGNITION A VS. RECOGNITION B	
		Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
(1)	7-response (g)*	.871 ± .011	.873 ± .011	.816 ± .015	.861 ± .111	.800 ± .016	.839 ± .013
(2)	7-response (n)†	.927 ± .006	.926 ± .006	.872 ± .012	.898 ± .009	.886 ± .010	.907 ± .008
(3)	5-response (g)	.907 ± .008	.910 ± .008	.860 ± .011	.903 ± .008	.864 ± .011	.902 ± .008
(4)	5-response (n)	.891 ± .009	.918 ± .007	.836 ± .013	.870 ± .010	.862 ± .011	.882 ± .010
(5)	3-response (g)	.838 ± .013	.848 ± .012	.797 ± .016	.875 ± .010	.837 ± .013	.858 ± .011
(6)	3-response (n)	.845 ± .014	.915 ± .007	.852 ± .012	.902 ± .008	.886 ± .010	.890 ± .009
(7)	2-response (g)	.859 ± .012	.865 ± .011	.735 ± .021	.806 ± .016	.745 ± .020	.864 ± .011
(8)	2-response (n)	.740 ± .018	.775 ± .016	.752 ± .018	.868 ± .010	.859 ± .011	.843 ± .012
(9)	True-false (g)	.804 ± .015	.839 ± .013	.675 ± .024	.801 ± .016	.641 ± .026	.780 ± .017
(10)	True-false (n)	.749 ± .018	.860 ± .011	.768 ± .017	.856 ± .011	.884 ± .009	.837 ± .013
(11)	Correlation of Recall A vs. Recall B					.950 ± .001 (100 items)	
(12)	Coefficient of reliability (Sum of Forms A and B)					.974 ± .001 (200 items)	
(13)	Averages of rows 5 to 10	.806	.850	.763	.851	.809	.845
(14)	Averages of rows 7 to 10	.788	.835	.732	.833	.782	.831

*(g) indicates the tests taken under instructions to guess.
†(n) indicates the tests taken under instructions not to guess.

It is to be noted that *all* recognition (multiple-response and true-false) tests were corrected for chance by the formula $R - \frac{W}{(n-1)}$, although five- and seven-response tests are never corrected in actual practice. Our chief interest therefore attaches to those tests having two or three responses. Rows (13) and (14) of Table 55 give the average correlations, both validities and reliabilities, for all of the two- and three- response tests and for all of the two-response tests, respectively.

The results are quite in harmony with Wood's results in the sense that validities are increased by the use of corrections for chance. They disagree with Wood, and all preceding investigations, in that correction for chance *increased* reliability. This is somewhat surprising, but much weight must be given to the fact that this is the most extensive investigation yet published, each correlation being based upon more than 200 cases and almost 2500 cases being used to find the reliability of the recall tests, Rows (11) and (12).

E. P. Wood was the next to publish results bearing on the question of correcting scores for chance.[1] Mrs. Wood used a lengthy criterion built up from seven separate measures. The reliability and validity of this criterion is hardly open to question. She gave to 147 college students in a course in government four tests as follows:

(*a*) A true-false "do not guess" test of 210 items.

(*b*) A 5-response "do not guess" test of 159 items.

(*c*) A simple recall "do not guess" test of 227 items.

(*d*) An old-type test (given two days prior to the three objective tests).

These tests did not cover "precisely" the same information; all were intended to be the best possible one-hour test on government. The items were largely problematical rather than informational. Her principal results follow:

[1]E. P. Wood, "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Psychology* Vol. XVIII (1927), pp. 18-25.

Test	Correlations With the Criterion		No. of Items
	Rights	$R-\frac{W}{(n-1)}$	
True-false	.748	.845	213
Five-response	.850	.860	159
Simple-recall	.880	...	227

As Mrs. Wood concludes, her results are quite in harmony with those of Ben D. Wood and of Ruch and DeGraff.

Effects of instructions concerning guessing. Table 60 presents the findings of DeGraff and Ruch[1] on the value of attempting to control the chance factors in recognition tests by the use of instructions for and against guessing.

TABLE 60

Effects of "Guess" and "Do Not Guess" Instructions on Test Scores (After DeGraff and Ruch)

Type of Test	Average Score on 200 Items		Differences
	(*a*) Rights	(*b*) $R-\frac{W}{(n-1)}$	(*a* − *b*)
Recall	55.4	....	....
7-Response (g)	89.6	74.1	15.5
7-Response (n)	82.1	73.1	8.0
Differences	7.5	1.0	
5-Response (g)	99.7	78.8	20.9
5-Response (n)	90.9	78.7	12.2
Differences	8.8	0.1	
3-Response (g)	118.8	80.1	38.7
3-Response (n)	102.3	78.0	24.3
Differences	16.5	2.1	
2-Response (g)	138.9	80.7	58.2
2-Response (n)	125.4	85.9	39.5
Differences	13.5	−5.2	
True-false (g)	127.2	59.3	67.9
True-false (n)	108.6	57.6	51.0
Differences	18.6	1.7	

[1] *Objective Examination Methods in the Social Studies*, p. 87.

Comparison of the differences of Column (*a*) of Table 60 shows clearly that perhaps fifty per cent of the guessing in multiple-choice tests can be eliminated by instructing pupils against guessing. This finding of itself proves nothing as to the merits of guessing or omitting unknown items. There is a sharp division of opinion on this point. McCall favors "guess" instructions, Wood insists upon "do not guess." Most authors say nothing about guessing in their test instructions (judging by standard tests). The author has always tended to agree with Wood, particularly in the light of the work of DeGraff and Mrs. Wood. McCall and others incline to think that the more items guessed at, the more adequate the correction formula will be in eliminating such effects. The real test is the effect of such instructions on the validity and reliability of the scores. Table 61 gives us just such an analysis of the results of the experiments of DeGraff and Ruch.

Table 61 shows fairly clearly that instructions against guessing increase the reliability somewhat, especially in comparisons with uncorrected scores (rights) when pupils are told to guess at all times. Table 62 presents a similar analysis of validity coefficients for the same investigation.

Table 62 indicates again the slight superiority of "do not guess" instructions, especially when correction of scores is practiced in connection with warnings against guessing.

It must be admitted that the evidence in favor of non-guessing at items (the answering of which appears to be sheer chance) is not very conclusive. We must leave this issue with the tentative recommendation against widespread guessing. The evidence for this conclusion, as we have seen, rests on three facts:

(1) Instructions against guessing lowered average scores (rights) about ten to fifteen per cent. This means that pupils tended to omit rather than to guess, although there is no means of knowing how much guessing was still present.

TABLE 61

RELIABILITY COEFFICIENTS FOR "GUESS" AND "DO NOT GUESS" DIRECTIONS FOR R AND $R-W$ SCORES
(After DeGraff and Ruch*)

TEST	"GUESS"			"DO NOT GUESS"			DIFFERENCES			
	1 Uncorrected	2 Corrected	3 Difference, 2−1	4 Uncorrected	5 Corrected	6 Difference, 5−4	4−1	4−2	5−1	5−2
Recall (.950).....										
7-Response.....	.800	.839	+.039	.886	.907	+.021	**+.086**	+.047	**+.107**	**+.066**
5-Response.....	.864	.902	+.038	.862	.882	+.020	−.002	−.040	+.018	−.020
3-Response.....	.837	.858	+.021	.886	.890	+.004	**+.049**	+.028	**+.053**	+.032
2-Response.....	.745	.864	**+.119**†	.859	.843	−.016	**+.114**	−.005	**+.098**	−.021
True-false.....	.641	.780	**+.139**	.885	.837	**−.048**	**+.241**	**+.105**	**+.196**	+.059
Average r.....	.777	.849	.....	.876	.872					

**Journal of Educational Psychology*, Vol. XVII (1926), p. 371.

†Values in bold-face type show all differences which are 3.0 or more times their probable errors, and hence are probably "significant" differences.

TABLE 62

VALIDITY COEFFICIENTS FOR "GUESS" AND "DO NOT GUESS" DIRECTIONS FOR R AND $R - W$ SCORES
(After DeGraff and Ruch*)

RECALL VS.	"GUESS"			"DO NOT GUESS"			DIFFERENCES			
	1 Uncorrected	2 Corrected	3 Difference, 2−1	4 Uncorrected	5 Corrected	6 Difference, 5−4	4−1	4−2	5−1	5−2
7-Response A	.871	.873	+.002	.927	.926	−.001	**+.056**	**+.054**	**+.055**	**+.053**
7-Response B	.816	.861	+.045	.872	.898	+.026	+.056	+.011	**+.082**	+.037
5-Response A	.907	.910	+.003	.891	.918	+.027	−.016	−.019	+.011	−.008
5-Response B	.860	.903	+.043	.836	.870	+.034	−.024	**−.067**	+.010	−.033
3-Response A	.838	.848	+.010	.845	.915	**+.070**	+.007	−.003	**+.077**	**+.067**
3-Response B	.797	.875	**+.078**	.852	.902	**+.050**	+.055	−.022	**+.105**	+.027
2-Response A	.859	.865	+.006	.740	.775	+.035	**−.119**	**−.125**	**−.084**	**−.090**
2-Response B	.735	.806	+.071	.752	.868	**+.116**	+.017	−.054	**+.133**	**+.062**
True-false A	.804	.839	+.035	.749	.860	**+.111**	−.055	**−.090**	**+.056**	+.021
True-false B	.675	.801	**+.126**	.768	.856	**+.088**	**+.093**	−.033	**+.181**	+.055
Average r	.815	.858	.....	.823	.890					
Proportion of "significant" differences (bold-face type) to total number of differences (10)	+	2:10	.....	.....	5:10	2:10	1:10	7:10	3:10
	−	0:10	.....	.....	0:10	1:10	3:10	1:10	1:10
	Both	2:10	.....	.....	5:10	3:10	4:10	8:10	4:10

Loc. cit., p. 371.

(2) Validities were slightly higher for "do not guess" directions, especially when scores were corrected.

(3) Reliabilities were somewhat higher for "do not guess" instructions, particularly when compared with uncorrected scores under instructions to guess.

Is the true-false test a two-response test? Tables 48, 50, 51, and 60 have raised definitely a question whether true-false and two-response multiple-choice tests are equally difficult. When the "same" items are administered in each of these two types, the true-false appears to be significantly the more difficult ("same" being defined in the sense shown by the items used below as an illustration). Granting, for the sake of discussion, that this is a true finding, what is the probable explanation?

One theory that has been frequently advanced, although not in explanation of the difference under discussion, is that false statements have a negative suggestion effect. If this is true, part or all of the difference may be accounted for. We shall see in the next chapter that the evidence for such a negative suggestion effect is not wholly conclusive; in fact, the evidence points to the fact that such an effect is of small consequence.

A second hypothesis advanced by the author[1] may be somewhat more promising. As a basis for discussion, the following sample items are quoted from the investigation by DeGraff and Ruch:

TWO-RESPONSE

1. Christopher Columbus discovered America in the year (1) 1498 (2) 1492

2. The first Pilgrims were brought to American shores in the ship named the (1) Mayflower (2) Half Moon

3. Eli Whitney is noted for his invention of the (1) spinning jenny (2) cotton gin

[1] *Journal of Educational Psychology*, Vol. XVII (1926), pp. 374-375.

4. Robert Fulton's contribution to civilization was the development of the (1) Atlantic cable (2) steamboat

5. The first permanent English settlement in America was (1) Plymouth (2) Jamestown

TRUE-FALSE

1. Christopher Columbus discovered America in the year 1492. .....

2. The first Pilgrims were brought to America in the ship named the Mayflower. .....

3. Eli Whitney is noted for his invention of the spinning jenny. .....

4. Robert Fulton's contribution to civilization was the development of the steamboat. .....

5. The first permanent English settlement in America was Plymouth. .....

Certain facts are obvious, a priori:

1. If a pupil is absolutely ignorant of an item, the chances of success are 50:50 on both true-false and two-response tests.

2. If he has a "fringe" of knowledge in either case, he will succeed more often than he fails.

3. If the chances of success are 50:50, the chance correction formula $(R-W)$ is equally valid in both types of tests.

4. If suggestibility enters into either type of test, it is equally divided between the two responses of the true-false test;[1] or at least approximately so.

5. If suggestibility, in the sense of a tendency to accept any statement suggested as true as being true, or vice versa, enters into the answering of true-false tests, there will be no *net* effects of an excess of statements marked either "true" or "false," if the test contains an equal number of true and false statements.

[1]Mathews, *Journal of Educational Psychology*, Vol. XVIII (1927), pp. 445-457, on the contrary, has brought forward evidence that in two-response tests the first stated alternative is selected 3.2% oftener than is the response which comes second, and also that the upper response is selected 33.8% oftener than the lower when the two are printed vertically. This issue will be discussed later. A student of the author has just completed a repetition of Mathews's study.

Returning now to the second hypothesis, let us divide any statement of an item into two parts: (1) the *critical statement*, and (2) the *completion*.

Thus, for item 2 the *critical statement* is: "The first Pilgrims were brought to American shores in the ship named the ". . .

The *completions* are: (true-false) Mayflower; (two-response) (1) Mayflower (2) Half Moon. Note that the true-false type does not mention at all the *Half Moon*.

If a pupil is ignorant of the name of the ship which brought the first Pilgrims to America, it does not necessarily imply that he is equally ignorant of the *Half Moon*. He may have, in fact, associations which couple the *Half Moon* with Henry Hudson. If so, he has a basis, indirect to be sure, of arriving at the correct answer *by elimination*. The chances are obviously not 50:50 here, although they might be if he met the statement in true-false form.

The point of the foregoing discussion is that both true-false and two-response tests present 50:50 chances for success when absolute ignorance is the case, but that the two-response is more open to successful answering than is the true-false when knowledge is present about either one of the two response words. This fact alone might, in the long run, make two-response tests somewhat less difficult.

The next section shows a method by which this problem may be attacked statistically.

The determination of the best method of scoring multiple-response tests through partial and multiple correlations.[1] Thurstone[2] seems to have suggested the selection of the best scoring method through the technique of multiple correlation. Starting with the general formula $S=R+CW$, in which S is the score, R the number of rights, W the number

[1]The student not familiar with partial and multiple correlation methods will probably not follow the full argument of this section.

[2]L. L. Thurstone, "A Scoring Method for Mental Tests," *Psychological Bulletin*, Vol. XVI (1919), pp. 235-240.

of wrongs, and C is a numerical constant which would determine the amount of deduction for errors, he derives the following formula for finding C:

$C=\frac{\sigma_R(r_{IR}\cdot r_{RW}\cdot r_{IW})}{\sigma_W(r_{IW}\cdot r_{RW}-r_{IR})}$, in which R, W, and C have the meaning stated above, and I means the *criterion* variable.

It should be noted at the outset that in case C takes on the value -1.00, this general formula becomes $S=R-W$, or the usual one. Moreover, it is just as convenient to employ the more familiar form of solution of a problem in multiple correlation, and especially so since Thurstone did not expand his derivations to include the fourth possible variable, viz., the number of omissions. Evidence to be presented here shows that a general formula of the type $S=C_1R+C_2W+C_3O$ is needed when instructions to omit items rather than to guess are used. (Thurstone apparently had in mind that all items would be attempted. At the time Thurstone wrote it was almost universal practice to instruct students to guess when in doubt.)

If evidence continues to accumulate to demonstrate that the value of the constant C in Thurstone's formula is not -1.00, it will be possible eventually to replace the a priori formula $S=R-\frac{W}{(n-1)}$ by empirical multiple-regression equations.

Several investigators, including Thurstone, have found that the values of C are not exactly the a priori ones (which are the ones commonly used when the $R-W$, $R-\frac{1}{2}W$, $R-\frac{1}{3}W$, etc., are used for two-, three-, and four-response tests, respectively).[1] Brinkley,[2] following the procedure of Thurstone, tried out a number of methods of scoring tests.

[1]Dr. T. L. Kelley, about 1921, mentioned to the author in the course of a private conversation that certain studies of his own had shown that $R-.9W$ was more defensible than $R-W$ for scoring true-false tests.

[2]"New-Type Examinations in the High School," pp. 94-97.

Before starting with a discussion of the results of Brinkley and others, it will be well to define the subscripts used throughout this section: (1) represents the criterion or dependent variable. (2) represents the number right. (3) represents the number wrong. (4) represents the number of omissions.

Table 63 gives an abstract of the experimental findings presented by Brinkley.

TABLE 63

SUMMARY OF BRINKLEY'S STUDY OF THE PROPER SCORING FORMULA FOR MULTIPLE-CHOICE TESTS

CORRELATIONS	TRUE-FALSE (31 Items)	3-RESPONSE (30 Items)	4-RESPONSE (30 Items)	5-RESPONSE (30 Items)
r_{12}	.784	.677	.667	.763
r_{13}	−.742	−.660	−.570	−.615
r_{23}	−.753	−.926	−.894	−.875
Best weighting for Wrongs (Thurstone's C)	−.81	−.53	+.18	+.22
Correlation with criterion when scored $R-\frac{W}{(n-1)}$	.82	.68	.67	.76

One other point should be noted about Brinkley's results, viz., that scoring by the formula $S=R-\frac{W}{(n-1)}$ gives better results than "number right" only in case of the true-false, and here the difference was but 0.04. Brinkley apparently did not attempt to find out whether the formula $S=R-.81W$ (using the value he found for C) would have been even better than $R-W$ for the true-false test. Brinkley's results suggest that the $R-W$ formula *over*-corrects true-false tests, i.e., penalizes unduly. A better formula would be $R-.8W$, or rights minus four-fifths the wrongs. We shall see later how well this agrees with later studies.

Two other studies may be mentioned, those of Mrs. E. P. Wood and of R. R. Foster and G. M. Ruch, curiously enough published simultaneously.[1]

Mrs. Wood used an elaborate criterion composed of seven separate measures pooled. The subject was a course in government, and 147 cases were used in finding the correlations. It is to be noted that Mrs. Wood used tests under instructions "not to guess," and hence her results show the influence of the fourth variable, omissions; in this respect the study is analogous to that of Foster and Ruch. Table 64 shows her results.

TABLE 64

SUMMARY OF E. P. WOOD'S RESULTS ON THE SCORING OF TESTS WHEN CHANCE IS INVOLVED

CORRELATIONS	TRUE-FALSE (213 Items)	5-RESPONSE (159 Items)	COMPLETION (227 Items)
r_{12}	.748	.850	.880
r_{13}	−.264	−.247	−.085
r_{14}	−.480	−.615	−.730
r_{23}	.145	−.146	−.052
r_{24}	−.864	−.792	−.840
r_{34}	−.607	−.459	−.459
$R_{1.234}$	.847	.861	.890
r_{12}	.748	.850	.880
Gains from use of wrongs and and omissions	.099	.011	.010

The last line of entries in Table 64 was computed by the author. It is to be noted that this use of wrongs and omissions is of no help in five-response and completion tests, but raises the validity about .10 in the case of the true-false test. Unfortunately Mrs. Wood did not publish her final

[1] E. P. Wood, "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Psychology*, Vol. XVIII (1927), pp. 18-25.
R. R. Foster and G. M. Ruch, "On Corrections for Chance in Multiple-Response Tests," *ibid.*, pp. 48-51.

regression equations. She did not compute the values for C in Thurstone's formula.

The study of Foster and Ruch differed somewhat from that of Mrs. Wood in several respects. In the first place (using the data of DeGraff and Ruch), the criterion was that of simple-recall scores (reliability .97) on the "same" items (200 in number) as used in all multiple-response tests employed. Second, the instructions were varied to include both "guess" and "do not guess" directions. In the third place, the numbers ran somewhat larger (221-281). Lastly, the subjects were elementary- and high-school pupils in history. Tables 65 to 69 give the results.

TABLE 65

CORRELATION COEFFICIENTS OF THE ZERO ORDER

	TRUE-FALSE	2-RESPONSE	3-RESPONSE	5-RESPONSE
INSTRUCTIONS TO GUESS				
r_{12}	.818 ± .014	.901 ± .008	.875 ± .010	.905 ± .007
r_{13}	−.769 ± .017	−.894 ± .009	−.757 ± .018	−.701 ± .022
r_{23}	−.627 ± .026	−.903 ± .008	−.675 ± .023	−.582 ± .029
INSTRUCTIONS NOT TO GUESS				
r_{12}	.784 ± .016	.768 ± .016	.887 ± .009	.897 ± .007
r_{13}	−.228 ± .039	−.381 ± .034	−.405 ± .037	−.364 ± .035
r_{14}	−.510 ± .030	−.466 ± .031	−.549 ± .031	−.583 ± .026
r_{23}	−.261 ± .038	−.038 ± .040	−.107 ± .044	−.117 ± .040
r_{24}	−.897 ± .008	−.841 ± .011	−.816 ± .014	−.811 ± .013
r_{34}	−.651 ± .023	−.495 ± .030	−.476 ± .034	−.477 ± .031

TABLE 66

COEFFICIENTS OF MULTIPLE CORRELATION

	TRUE-FALSE	2-RESPONSE	3-RESPONSE	5-RESPONSE
$R_{1 \cdot 23}$ (guess)	.882	.920	.903	.930
$R_{1 \cdot 234}$ (do not guess)	.903	.846	.940	.935

TABLE 67

REGRESSION EQUATIONS FOR THE RAW OR GROSS SCORES

INSTRUCTIONS TO GUESS	
1. True-false	$\overline{X}_1 = .867X_2 - .772X_3 - 2.358$
2. 2-response	$\overline{X}_1 = .762X_2 - .693X_3 - 11.553$
3. 3-response	$\overline{X}_1 = .716X_2 - .384X_3 - 4.853$
4. 5-response	$\overline{X}_1 = .666X_2 - .268X_3 + 7.439$
INSTRUCTIONS NOT TO GUESS	
1. True-false	$\overline{X}_1 = .785X_2 - .806X_3 - .059X_4 + 14.709$
2. 2-response	$\overline{X}_1 = .511X_2 - 1.114X_3 - .433X_4 + 61.032$
3. 3-response	$\overline{X}_1 = .632X_2 - .505X_3 - .114X_4 + 19.916$
4. 5-response	$\overline{X}_1 = .337X_2 - .720X_3 - .409X_4 + 85.362$

TABLE 68

REGRESSION EQUATIONS USING STANDARD MEASURES[1]

INSTRUCTIONS TO GUESS	
1. True-false	$\overline{Z}_1 = .553Z_2 - .422Z_3$
2. 2-response	$\overline{Z}_1 = .508Z_2 - .435Z_3$
3. 3-response	$\overline{Z}_1 = .668Z_2 - .305Z_3$
4. 5-response	$\overline{Z}_1 = .751Z_2 - .263Z_3$
INSTRUCTIONS NOT TO GUESS	
1. True-false	$\overline{Z}_1 = .841Z_2 - .501Z_3 - .082Z_4$
2. 2-response	$\overline{Z}_1 = .421Z_2 - .557Z_3 - .401Z_4$
3. 3-response	$\overline{Z}_1 = .738Z_2 - .390Z_3 - .149Z_4$
4. 5-response	$\overline{Z}_1 = .397Z_2 - .573Z_3 - .541Z_4$

[1] Z is defined as $\frac{X-M}{\sigma} = \frac{x}{\sigma}$.

The results of Wood and of Foster and Ruch show that there is some theoretical advantage in taking into account both wrongs and omissions in scoring true-false, two-response, and possibly three-response. Coupled with the preceding evidence to the effect that it is somewhat better to

TABLE 69

IMPROVEMENT OF $R_{1\cdot23}$ OR $R_{1\cdot234}$ OVER r_{12} IN THE RESULTS OF FOSTER AND RUCH

Correlation	True-False	2-Response	3-Response	5-Response
Instructions to Guess				
$R_{1\cdot23}$	.882	.920	.903	.930
r_{12}	.818	.901	.875	.905
Gain	.064	.019	.028	.025
Instructions Not to Guess				
$R_{1\cdot234}$	.903	.846	.940	.935
r_{12}	.784	.768	.887	.897
Gain	.119	.078	.053	.038

use instructions against guessing, it may be tentatively concluded that Thurstone's proposals (or the equivalent multiple-regression equation method) might well be employed in studies requiring a high degree of accuracy. It can hardly be defended that such refinements are needed in actual classroom tests. It is clear that even the omissions have some significance in arriving at the best scoring formula. On the other hand, the teacher employing tests of 100 to 250 items may safely score her tests simply "number right" or perhaps $R-\frac{W}{(n-1)}$ in the case of three responses or fewer, especially when pupils are instructed against wild guessing.

Thurstone's formula for obtaining the value for C (the best weighting of wrongs) cannot be applied to the data of E. P. Wood or the results of Foster and Ruch under "do not guess" instructions because it makes no allowance for the fourth variable in the situation, viz., errors. A formula which would also allow for the inclusion of errors may be derived quite readily. In an absence of such a formula, the four-variable multiple-regression equation may be applied,

particularly if standard measures (sigma values) are employed.[1]

Table 70 shows the values for C found by Brinkley and by Foster and Ruch, as well as those suggested by the formula $S = R - \frac{W}{(n-1)}$.

TABLE 70

VALUES OF C FOR THE RESULTS OF BRINKLEY AND FOSTER AND RUCH

	TYPE OF TEST			
	True-false	2-Response	3-Response	5-Response
Values of C as found by Brinkley	−.81	.....	−.53	.22
Values of C as found for Foster and Ruch data	−.89	−.91	−.54	−.40
Values of C indicated by $R - \frac{W}{(n-1)}$ formula	−1.00	−.100	−.50	−.25

The foregoing table shows far from perfect agreements of calculated and expected values (as obtained by the conventional scoring formula). Only in the case of the three-response is there reasonably close agreement. This table shows the need for extended study of the issue in question.

Such data as are at hand suggest the possibility that the $R - W$ formula over-corrects by at least ten per cent, possibly by as much as fifteen to twenty per cent in two-response and true-false tests. In fact, there is some evidence of over-correction (undue penalization) in all multiple-response tests.

On the other hand, even an over-correction of fifteen to twenty per cent is not of very great practical significance in

[1]This implies assuming rectilinearity of regression, of course. It should be noted that Foster and Ruch found many correlations (especially when "do not guess" directions were employed) which were markedly curvilinear. Wood did not comment on the degree of rectilinearity of her data.

informal classroom testing, although the available evidence suggests that careful experimental studies may well afford to employ either the technique of Thurstone or the more familiar (but equivalent) procedure of Foster and Ruch.

In conclusion, it may not be out of place to suggest that studies along the direction suggested by Thurstone, Brinkley, Wood, and Foster and Ruch be multiplied until generalizations may be made leading toward a more final solution of issues discussed here.

Some proposals for modified true-false tests. There have been numerous proposals looking toward a betterment of the true-false test. These have been directed chiefly at the elimination of chance errors, although some have been concerned with the matter of avoiding negative suggestion effects through the use of the interrogative statement followed by "yes-no" or "right-wrong."

The "true-false-doubtful" and the "yes-no—didn't-say" illustrate another line of thinking. (See Chapter VIII for examples.)

Recently Christensen[1] and Greene[2] have suggested interesting variations of the true-false. The former proposes using the "same" item in both true-false and multiple-choice form in the same test, i.e., each item occuring twice. To earn one unit of score, *both* items of such pairs must be correct. He found the reliabilities to increase markedly in such arrangements. Thus, for 100 true-false and multiple-choice items in general science, Christensen obtained the following reliabilities:

True-false alone ($R - W$ scores)	.67
True-false alone (R scores)	.76
Multiple-choice alone (R scores)	.80
True-false and multiple-choice combinations (Christensen's plan)	.90

[1] *Journal of Educational Research*, Vol. XIV (1926), pp. 370-374.
[2] *Ibid.*, Vol. XVII (1928), pp. 102-107.

Similar gains were shown for validity when the Van Wagenen *General Science Reading Scale A* was used as a criterion.

He shows further that to obtain a reliability of .90 (that of the combined plan) would require tests of these lengths:

True-false ($R-W$)	450 items
True-false (R)	288 items
Multiple-response	200 items

Moreover, the combination plan would save time. Against this undeniable advantage in reliability are certain rather serious disadvantages: (*a*) the scoring is somewhat complicated, and (*b*) 200 items give no wider sampling than do 100 items of either true-false or multiple-choice. To double the sampling might well prove to be more valid than to double up on half that number of items. This point should be investigated before accepting this suggestion.

Greene's suggestion is slightly more workable since he has simplified the problem of scoring and tabulating. This writer would give each item both as a true and as a false statement. Items one to fifty present the statements, and items fifty-one to one hundred repeat the former, making the true statements false, and vice versa. As before, one point credit is given for each *pair* correct. Greene's evidence does not appear very conclusive, as he points out. Moreover, Greene comments on the fact that 100 *different* items may be superior to fifty paired items. His suggestion is made more in the effort to encourage study of methods of bettering true-false tests than as a practical suggestion for immediate adoption.

Another proposal for the modification of the true-false test is that of McClusky and Curtis. The suggestion of these writers is that pupils mark all true statements with a "T" but "correct the statements you consider to be false by *substituting* an appropriate word or phrase for some

word or phrase in the original statement so as to make it true. . . . . No credit will be given for false statements which are corrected merely by the insertion of the word 'not'."[1]

The proposed method requires roughly double the working time of the regular true-false test in the seventh grade, about forty per cent more time in high school, and about one-sixth more time in college classes. The average scores on the modified test are somewhat smaller, and more time is needed for scoring.

Except in one case, the modified form showed from .15 to .20 higher reliability coefficients. In general, the reliability coefficients under the modified form are little higher than would have been the case had the same amount of time been used for the regular true-false test, i.e., had equal working times been compared using the Spearman-Brown prophecy formula. This fact tends to discount the value of this proposed innovation. In view of the extra labor, the net gain of this method would be very small indeed.

Chapter summary. The following statements summarize the findings presented in this chapter, although many of the statements made must be taken with great caution until further experimental evidence is at hand:

1. There have been two general approaches to the problem of the correction of test scores for chance effects, viz., a priori or mathematical discussions based upon theorems of probability and experimental studies. The latter are probably more trustworthy.

2. Many criticisms of the true-false test (especially) have been based upon a misunderstanding of the significance of probability.

3. When there are no omitted items, the score may be computed either as "number right" or "rights minus

[1]H. Y. McClusky and F. D. Curtis, "A Modified Form of the True-False Test," *Journal of Educational Research*, Vol. XIV (1926), pp. 213-224.

wrongs," since such scores correlate perfectly (Holzinger's proof).

4. Wrong answers are of at least two kinds: (*a*) answers guessed at, but incorrect; (*b*) answers not guessed at, but answered in good faith, although products of straight misinformation.

5. The binomial theorem will give the mathematical expectancies of right and wrong responses, *provided that all wrong answers are pure guesses.* It cannot cover the issue of misinformation and consequent wrong responses.

6. Whether correction for chance lowers or raises the reliability is still debatable. Most of the studies point toward a lowering, although one of the most recent and extensive points toward the opposite effect.

7. Regardless of slight effects upon reliability, the use of corrections for chance should be decided upon the run of the evidence bearing on the effects of such corrections upon *validity.*

8. The evidence is almost without disagreement to the effect that correction for chance increases the validity of test scores, especially when true-false, two-response, and three-response tests are concerned.

9. The teacher who does not wish to trouble to correct test scores for chance may avoid this labor by making her tests ten to fifteen per cent longer than originally planned and thus eliminate the need for correction.

10. The available evidence suggests that both more valid and reliable scores are to be obtained by instructing pupils to omit items where the answering is nothing more than a sheer guess, i.e., by using "do not guess" instructions. This does not mean that pupils should not attempt items about which they have definite "hunches" or "fringes of knowledge."

11. The two investigations (E. P. Wood and DeGraff and Ruch) which studied the combined question of instruc-

tions about guessing and corrections for chance suggest that the best practice is (*a*) instructions against widespread guessing, combined with (*b*) corrections for chance.

12. Some argumentative evidence was introduced to the effect that the true-false test differs considerably in its psychology from the two-response test proper, in that the latter suggests more facts which might form the basis for the answering of the item.

13. Thurstone has provided a convenient formula for weighting wrong responses when "guess" instructions are employed.

14. The method of Thurstone and the multiple regression equation method of Foster and Ruch suggest that the $R - W$ formula over-corrects from ten to twenty per cent in the case of true-false and two-response tests (as applied to the data of Brinkley and of Foster and Ruch).

15. There is some evidence that the general formula $S = R - \frac{W}{(n-1)}$ over-penalizes in all multiple-response tests, but much work is yet to be done on this problem.

CHAPTER XIII

THE NEGATIVE AND OTHER SUGGESTION EFFECTS IN THE TRUE-FALSE TEST

The issue stated. Modern psychology of teaching emphasizes the danger of *practice in error*, i.e., the exercising of wrong neural connections. In teaching spelling or arithmetic, for example, the teacher makes every effort to catch errors before constant repetition stamps in such wrong reactions. It must be admitted that associations are built up in exactly the same manner, whether the ideas associated are right or wrong.

Many critics have held the false statement of the true-false examination to be dangerous pedagogy in that it tends to implant misinformation in the mind of the pupil. Since about fifty per cent of the statements in such a test represent falsities, it is evident that great opportunity exists for fixing error in the minds of pupils reading such false statements. It is perhaps more than a little curious that such critics have never seemed to reckon with the fact that the true statements might tend to fix truths in the minds of pupils.

The question at issue has usually been known by the name of *negative suggestion.*

As has previously been mentioned, some authorities think that the stating of true-false items in question form will avoid this danger. Thus:

Is milk white?	*Yes*	*No*
Do all plants have flowers?	*Yes*	*No*

So far as the author knows, the believers in the negative suggestion effect have discontinued their psychologizing at the point where they discovered the possibility of such negative learning.

There is, as a matter of fact, a great deal known in modern psychology which bears on this issue. For example, there is the psychology of the *mental set* or *idea-in-mind*. Undoubtedly the danger of negative learning is greatly dependent upon the attitude of mind of the pupil taking such a test.

We must grant at once that if a page of spelling words, of which approximately fifty per cent were misspelled, were placed before a pupil for study, there would be much practice in error. The sanctity of the textbook would see to that. On the other hand, the attitude of a pupil toward a true-false test is not the uncritical, passive, conscious set which we have described in the case of the page of spelling; but on the contrary, the directions to the test and past experience alike cause him to assume a critical, challenging, and thoughtful attitude. He knows in advance that about half of the statements are untruths. His task is to find which are which.

Moreover, there is some logic in assuming that pupils should be exposed to situations calling for discrimination between truth and error in exactly the same way that we have come to hold that real moral and ethical character cannot be attained through "running away from sin without a battle."

Pages might be written on both sides of this argument. It is better to turn to the meager experimental evidence which exists. It is to be admitted that this evidence is insufficient in quantity. On the other hand, if the negative suggestion effect is one-half or even one-fourth as great as some critics suppose, it would bob up alarmingly in even the crudest experimentation.

Ballard's investigation. The first study on the suggestion effects of true-false tests seems to be that of Ballard.[1] He gave true-false tests in geography and history to classes of thirteen-year-old boys. Later the same questions were

[1]P. B. Ballard, *The New Examiner*, (London: Hodder and Stoughton, 1924), pp. 96-98.

presented in short-answer form and discussed briefly in class. Still later the same questions were given in recall form. Some of his results follow:

Subject	Gains From Initial True-False to Final Recall	
	"Trues"	"Falses"
Geography	17%	30%
History (I)	7%	73%
History (II)	14%	64%

From this Ballard concludes (p. 96) ". . . . . it will be found, I think, that children will learn more from the true-false test than from any other type of examination, and it is a curious fact that they will learn more from the false items than from the true."

Ballard does not state how many cases were used, but all three classes showed uniformly that the gains on the false items of the (original) test were much larger than on the "trues." The significance of these results is not altogether clear; so we should examine certain further studies before attempting to draw conclusions.

The study of Remmers and Remmers.[1] These authors chose a selection from the "Customs of the Germans." On this passage they built 121 true-false and 121 simple completion (recall) statements, each statement being made in both true-false and recall types. Two groups of college students were equated by intelligence-test scores. These were designated as Groups I and II. On the first day of the experiment both groups were given the selection to study and were told that they would have a test over it at the next meeting of the class. At the next meeting Group I was given the true-false test, and Group II the recall test. So far as the students knew, this ended the experiment.

[1]H. H. and E. M. Remmers, "The Negative Suggestion Effect of True-False Examination Questions," *Journal of Educational Psychology*, Vol. XVII (1926), pp. 52-56.

About a month later and without warning the tests were repeated, except that Group I was now given the recall and Group II the true-false. To quote the authors (p. 54), "The logic on which we shall base our conclusions is as follows: If the average score of Group I be equal to or greater than that of Group II, it follows that the taking of a true-false test can have no deleterious effect upon the formation of correct associations." Table 71 presents their results.

TABLE 71

RESULTS OF REMMERS AND REMMERS STUDY OF THE NEGATIVE SUGGESTION EFFECT OF TRUE-FALSE TESTS

TEST MATERIAL	AVERAGES		DIFFERENCE OF AVERAGES	P. E. DIFF.	APPROXIMATE CHANCES OF SIGNIFICANT DIFFERENCES
	Group I	Group II			
True-false	71.16	67.88	3.28	3.35	1 to 1
Recall.....	106.18	105.00	1.18	1.44	1 to 1
Sum of T-F and Recall.....	177.34	172.88	4.46	3.37	1.6 to 1

If Group I, which took the true-false test the next day after reading the passage, should have gathered a number of false impressions, it should have showed many wrong answers after the month interval, and consequently a lower score when the recall test was taken. As a matter of fact, Group I did slightly better after four months than did Group II at the outset, although this difference has no statistical significance. The very slight effect noted was a positive rather than a negative suggestion effect.

The investigation of Roberts and Ruch. A somewhat more extensive investigation is that of Roberts and the author.[1] Two different types of experiments were carried out by Miss Roberts, as follows:

I. On a given day about half of the students of each of a number of high-school science classes took a completion

[1]H. M. Roberts and G. M. Ruch, "The Negative Suggestion Effect of True-False Tests," *Journal of Educational Research* (Sept., 1928), pp. 112-116.

test, and the other half took a true-false test on the "same" items. Two weeks later all took the completion test (half, of course, for the second time). These results were studied by the method of averages as employed by Remmers and Remmers.

The groups were later reversed with a second set of materials covering a different topic.

II. A completion test was followed immediately by a true-false test containing the "same" items. At intervals from one day to one month, the completion test was administered to all. The two completion tests were then checked, item by item, to see the extent to which the second answers differed from the first.

The subjects covered were biology, botany, physics, and chemistry. The tests contained from thirty-four to seventy-three items. All tests were understood by the pupils to bear on their term standing. In all possible respects this investigation was made to be a part of regular classroom routine.

There were some disagreements in the results of the experiments of the first type in different classes. The botany examinations showed no evidence of negative suggestion effects (groups of about thirty students). The chemistry examinations (three sections of about twenty each) gave a statistically significant, but small, evidence of such undesirable effects.

The experiments of the second type are far more significant, as the answers on each individual test item were followed in detail from one examination to the next. A total of 235 students was used in this series of testings, making 705 test papers in all. In addition, 128 students took the two recall tests but not the true-false as a check on results. Table 72 summarizes certain of these results.

A few statements summarizing Miss Roberts's results follow. It is to be noted that conclusions are never based upon differences less than three times their standard errors, in accord with common statistical dicta. Some of the fol-

TABLE 72

DIFFERENCES BETWEEN THE AVERAGE SCORES ON THE INITIAL COMPLETION TESTS AND THE SAME TESTS REPEATED AT INTERVALS OF FROM ONE TO THIRTY-FOUR DAYS WITH TRUE-FALSE TESTS COVERING THE "SAME" ITEMS INTERVENING, AND WHEN NO TRUE-FALSE TEST INTERVENED

Section	Interval in Days	Difference in Averages*	S. D. of Difference	Difference / S. D. (Diff.)	Nos.
I. SHORT INTERVALS					
Biology A	1	8.2	0.6	14.2	21
Chemistry IA	3	4.1	0.8	5.0	20
Chemistry SB	3	1.8	0.4	5.0	26
Chemistry SA'	4	2.1	0.5	4.1	24
Physics B	5	3.1	0.4	7.3	24
Total					115
II. LONGER INTERVALS					
Chemistry NA	12	2.5	0.7	3.7	19
Biology B	15	0.4	1.1	0.4	17
Chemistry IB	20	2.2	0.7	3.0	21
Chemistry NB	31	3.0	0.7	4.2	20
Physics C	33	0.8	0.8	1.0	22
Physics A	34	−0.4	0.7	−0.5	21
Total					120
III. LONGER INTERVALS, NO INTERVENING TRUE-FALSE TEST					
Botany A, Exam. 2	14	1.9	0.8	2.3	33
Botany B, Exam. 1	15	3.2	1.1	2.9	28
Chemistry A'	14	0.0	0.6	0.1	22
Chemistry A''	14	−1.3	0.8	−1.6	21
Chemistry B	14	1.6	0.6	2.4	24
Total					128

*Positive values indicate that the average on the second test was higher than on the first, and vice versa for negative values.

lowing conclusions are based upon Table 72 and some on data not reproduced here.

1. As a result of the true-false test, roughly, one answer in twenty was changed to the "identical" wrong (the one exposed in the false statements), when chance was allowed for.

This refers to answers originally correct on the initial completion test.

2. About one answer in seven was changed to an "identical" wrong as exposed in the false statements when the question was omitted on the original test.

3. The longer the interval the fewer the answers accepting wrong statements, thus suggesting that negative suggestion effects are relatively temporary. (The intervals ranged from one to thirty-four days.) After a month very few false impressions persisted.

4. The number of changes to "identical" true responses outweighed the negative effects. Thus the *net* effect of the true-false was a positive suggestion phenomenon. (The differences in this case ranged from four to fourteen times their standard errors.)

Summary. All in all, Miss Roberts's work is in substantial harmony with the results of Remmers and Remmers. This may be the more significant since Miss Roberts worked on the high-school level and Remmers and Remmers dealt with college students.

The results of the two studies are alike in that both showed the *net* suggestion effect to be positive and not negative. Miss Roberts's differences were, however, very much more significant statistically than were those of the earlier study. The principal difference is that Miss Roberts found slightly more negative suggestion operating, although five per cent seems a reasonable allowance.

Much more work is needed before these conclusions may be accepted as final. It is probably fair to suggest that a priori arguments, reasonable as they may appear to be, be discounted until the advocates of negative suggestion can bring forward empirical disproof of these two studies. It appears almost certain that the negative suggestion effect

of the true-false test has been over-rated in the minds of some critics.

The disparity between experimental results and theoretical expectations may lie in the attitude of mind, the mental set, assumed by most pupils in taking true-false tests. The critical faculties are ordinarily more active in the true-false situation than they are in such activities as reading a lesson from a text, reciting orally, or writing a discussional examination.

Do pupils tend to respond "True" oftener than "False" on true-false tests? This question, although not closely related to the question of the negative suggestion effect, has some general interest for us in these days when the true-false test "hangs in the balance."

Fritz[1] and others have claimed that pupils mark more statements "true" than they do "false," this author reporting the ratio to be about 62:38. It should be noted that the test employed used technical material quite unknown to the students. However, almost identical results were obtained with materials studied as regular class exercise.

The author has carried on a number of studies similar to that of Fritz. In one such, for 164 students taking an examination in educational psychology, the ratio of responses given as "true" to those marked "false" was but 52:48. There is therefore little agreement between these two studies.

In another unpublished study by two students of the author (D. D. Durrell and C. L. Cushman), seventy-four students taking a true-false test in psychology were asked to indicate whether their responses were (1) matters of absolute certainty, (2) matters of uncertainty, or (3) pure guessing. The percentages of errors (based upon the number of at-

[1]M. F. Fritz, "Guessing in a True-False Test," *Journal of Educational Psychology*, Vol. XVIII (1927), pp. 558-561.

tempts) were, respectively, 20.1 per cent, 33.6 per cent, and 48.8 per cent. These students were mature persons and likely to follow instructions rather closely. The results indicate that pure guessing, when it occurs, is roughly a 50:50 situation.

Rutledge[1] found that when the "same" items were stated in both true and false forms, they were equally difficult. This finding is not necessarily opposed to the common belief and finding that false statements are more difficult than true statements. Taking true and false statements as they come, it is entirely possible that the maker of the test should cause the false statements to be the more difficult. Weidemann, Rutledge, and others have shown that false statements tend to phrasings that are confusing by the use of double negatives, "trick" wordings, etc.

Further comment on these issues is withheld through a conviction that much more crucial experimentation is needed before drawing working conclusions.

The suggestion effect of position of printed response words. Mathews[2] has recently published evidence that, in two-response tests, the left-hand response is selected 3.2 per cent oftener than is the right-hand alternative, and that the upper alternative is chosen 33.8 per cent oftener than the lower.

H. W. Meyer is at present, under the author's direction, repeating Mathews's work. Meyer's results will be published later.

There is one proposal in Mathews's conclusions that needs comment, viz., that his solution calls for the alternation of the response words, *true* and *false*, from one item to the next. This arrangement is illustrated at the top of the next page.

[1]R. E. Rutledge, "The True-False Examination in Elementary Psychology, with Suggestions for its Improvement," Unpublished Ph. D. Thesis, University of California, (1927).
[2]*Journal of Educational Psychology*, Vol. XVIII (1927), pp. 445-457.

1. Starch is a carbohydrate.	True	False
2. Pepsin is an enzyme of the saliva.	False	True
3. The gastric juice is alkaline in reaction.	True	False
4. The gastric juice acts chiefly on proteins.	False	True
Etc.		

Such an arrangement is very likely to be sufficiently confusing to destroy any possible gains from the rotation of response words. Moreover, the author is unable to see that Mathews's finding, if verified, has a great deal of significance. The argument runs as follows:

1. It is unlikely (and Mathews did not attempt to prove) that pupils who actually know a response will be mislead by such a fact as mere position of response. (Mathews, in fact, shows definitely that this tendency to response to position increases as the items became more difficult.)

2. Assume that in all pure guesses the upper response is chosen 33.8 per cent oftener, and the left is selected 3.2 per cent oftener (or even *always*). The $R-W$ formula will take care of the situation, provided the test is constructed so that right answers occur equally often in upper vs. lower or left-hand vs. right-hand positions. Response by position is then equivalent to response by chance. There should be no net gain or loss by such an extraneous method of responding. It should be noted that in scoring the Army Alpha tests during the late war there were many cases where soldiers responded without exception to one or the other of the words "true" or "false." Such papers were arbitrarily marked zero on the assumption that $R-W$ scoring would yield zero if equal numbers of true and false statements occurred. If fewer than 100 per cent were responded to systematically by either "true" or "false," the effect is the same.

Chapter summary. There is little or nothing in the discussions of this chapter that dare be taken as proved. The following statements are highly tentative:

1. The negative suggestion effect of false statements in true-false tests is probably much smaller than is sometimes assumed.

2. The small amount of negative suggestion which has thus far been shown for true-false tests seems to be fully offset by net positive teaching effects.

3. It is not established that students when in doubt mark statements "true" oftener than they do "false," the evidence for this finding being quite inconclusive.

4. Rutledge claims that the identical items are equally difficult when stated as true and as false statements.

5. It appears that position of printed response words affects the answering, especially when the alternatives appear one above the other. Upper and left-hand responses are more frequently selected. This finding, if true, may have little practical significance.

CHAPTER XIV

EXAMINATIONS, MARKS, AND MARKING SYSTEMS

Classification of marking systems. There are doubtless many hundreds of different marking systems in use in the United States if we include all the minor variations of a few common types. In a survey of 281 Illinois high schools Odell[1] found nearly a hundred different plans in use. In spite of such diversity, almost all marking systems reduce to two general categories when we examine their fundamental logic and assumptions, viz.,

1. Systems based upon *absolute* (and usually subjective) standards, the most familiar example being the 100 per cent (or 100-point) scale.

2. Systems based upon *relative* values, ranks, or the normal curve.

The former are probably more common. Odell found them to be so in the ratio of about 3 to 1.

In speaking of the common 100-point[2] or percentage scale as an *absolute* scale, the reference is to the existence of a standard in the mind of the marker and not to the use of numbers per se as marks. Numbers or letters are equally defensible since either must be defined before they can be held to take on meaning. The only possible advantage of the use of letters is that this practice may tend to discourage marking plans which call for more grades or degrees of achievement than are distinguishable by the human judg-

[1]C. W. Odell, "High School Marking Systems," *School Review*, Vol. XXXIII (1925), pp. 346-354.

[2]We shall refer to this plan as the "100-point scale" in spite of the fact that there are really 101 possible marks from zero to one hundred, inclusive.

ment, aided even by the best objective measures which have yet been devised. The use of a numerical series like 1, 2, . . . 5 or 1, 2, . . . 7 is fully as good as letter series of 5 or 7 letters since 5 or 7 levels of ability are probably distinguishable with a reasonable degree of accuracy. To use the 100-point scale for such purposes may be objected to as giving resulting marks an appearance of accuracy which is chiefly spurious. The objection is precisely that of stating that a pupil's weight is 78.89 pounds when measurements taken five or ten hours apart might show variations of a pound or more. In other words, to attempt to mark one pupil 78 and another 79, while open to no theoretical objection, is to be severely criticized in the light of experimental findings to the effect that from three to ten (if as many as ten) degrees of ability represent the maximum discrimination in judging human nature.

THE PERCENTAGE GRADING PLAN

The percentage grading plan analyzed and criticized. The strongest argument for the 100-point scale in grading is that this practice is rather generally familiar and understood. In reality the greatest weakness of this system is that its sheer familiarity leads to an uncritical acceptance without conscious regard to the inherent assumptions. To employ a scale of marks beginning with 0 and ending with 100 (101 marks in all), implies either (*a*) that the user actually thinks that he is distinguishing 101 levels of pupil accomplishment, or (*b*) that he realizes the spuriousness of such an assumption, but through inertia or callousness to the accepted rules of science to the effect that results should not be stated with *apparent* accuracy greater than the real accuracy, he continues to use such a misleading scale of marks.

One further aspect of the 100-point scale is worthy of comment, viz., the use of a passing mark (usually between 65 and 80, most often at 70 or 75). Such a *fixed* passing

mark is another proof of the absolutism of such scales. It takes no account of the very important fact that the number of pupils reaching, exceeding, or falling short of 70 (or some other passing mark) is a function of the difficulty of examinations, subjects, etc., or of personal "standards" of passing work. These criticisms, again, are based upon practical rather than theoretical considerations.

We may now turn to a somewhat more detailed criticism of the percentage (100-point) scale of markings.[1]

1. The stated *zero* and *100* points of the 100-point scale are arbitrary, undefined, and inflexible. They are not defined in terms of any true scale of educational abilities. The true zero and the true 100 on a scale of absolute achievement will often fall long distances above or below such arbitrary limits, were true units of educational measurement at hand. The only thinkable zero point is that of "just not any ability" for a given school subject. What a *true* "100" means is conjectural. It means, a priori, "perfection." A more rational meaning is probably "as good as can be expected when all factors of the situation are taken into account." This throws the meaning of "100" quite upon a basis of subjective judgment, and gives it as many different meanings as there are teachers employing the 100-point scale.

2. If the zero and 100 points are fixed quantities, all intermediate points are, by the same token, fixed, and presumably the increments between successive integral values are equal. Thus, a pupil earning 71 in a school subject is exactly as much superior to one earning 70 as is one earning 98 to one receiving 97. As every one admits, such a strict interpretation is nonsense. Yet such are the implications of the use of the scale.

[1]There are some thinkers who object to branding the 100-point scale as a *percentage* scale. A little thought will show that such objections are quite pointless. If we assume a scale of marks beginning with 0 and ending with 100, it is entirely defensible to brand such a series "per cents." They are in effect such from mathematical considerations. The issue is something like that of the question whether Shakespeare wrote the plays commonly ascribed to him or whether these were written by another man of the same name.

3. As has been mentioned, the 100-point scale, if taken literally, implies 101 distinguishable differences in accomplishment.

The three foregoing criticisms are based upon a strict interpretation of the theoretical implications of the percentage scale. To these we must add two other criticisms which follow from the practical considerations in the use of the scale.

4. The use of an arbitrary passing mark (70 or some other value) is without adequate defense, and in its actual operation results in throwing the distribution of actual marks given into a skew distribution quite at variance with the probable facts. Judging from a mass of accumulated evidence, pupils of any school grade distribute themselves approximately normally, i.e., in rough accordance with the curve of chance, which in turn seems to hold reasonably well for most phenomena of pure biological and psychological variation. If some mark in the general vicinity of 70 is taken as "passing," the vast majority of pupils will be marked between 70 and 100. The average will ordinarily fall in the eighties of such a scale. Assume for purposes of discussion that the average will fall at or near 85, a reasonable approximation to the facts. This means that superaverage pupils distribute themselves along a 15-point scale (between 86 and 100), and sub-average pupils arrange themselves along a 85-point scale (between 0 and 84). Such a condition is, to say the least, illogical. As a matter of fact, we have been so disturbed by such a lop-sided scheme that we have consciously or unconsciously attempted to better conditions by the very simple, and quite illogical, expedient of ignoring the lower 50 points of the 100-point scale, except in rare instances. Examination of marks as actually given from teacher to teacher and from school to school is the best possible evidence that the 100-point or percentage scale does not work in practice.

If we are to grade upon a basis of 0 to 100, there is one and only one logical average mark, viz., 50. Otherwise we tend to force our marks into a skew distribution which is unlikely to represent the facts. To find any sure basis for a passing mark (which, by our logic, must lie between 0 and 50) is a quite impossible task unless we resort to arbitrary definition and selection of some point, as, say, 25 or 30.

There is no intention of arguing that marks should be distributed exactly normally. On the other hand, a distribution like that shown in Figure 12 is much less reasonable than one like Figure 13.

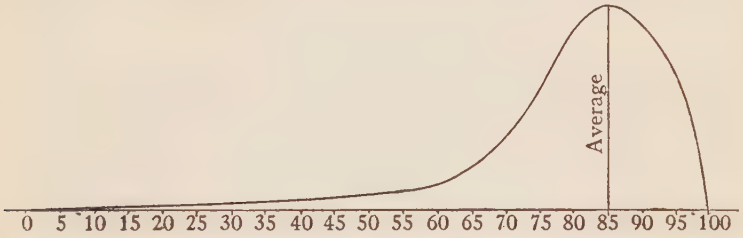


FIG. 12.—Showing the skewness and general form of the distributions of grades usually observed when the 100-point scale is applied.

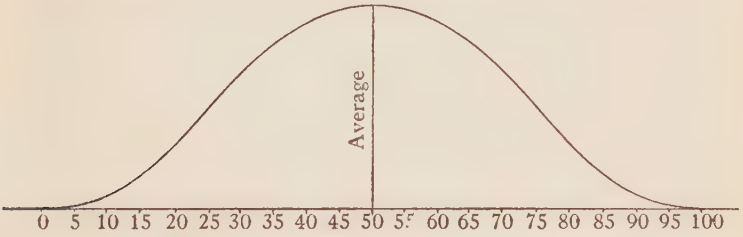


FIG. 13.—Showing a more rational use of the 100-point grading plan, and one more in harmony with the known facts about individual differences.

5. The fifth objection which we shall mention about the 100-point grading scheme is that a large body of experimental evidence points to the fact that from but five to seven levels of ability are ordinarily recognizable by teachers

in marking pupils. This is again a practical objection and one resting upon empirical evidence. The difference between an "85" and an "86" is a difference at least five times as fine as the human judgment can ordinarily distinguish.

Before leaving this treatment of the 100-point or percentage system, it is well to extend the argument to cover such systems as the following:

A equals 95 to 100.
B equals 85 to 94.
C equals 75 to 84.
D equals 70 to 74.
E equals below 70 (failure, condition, etc.)

In such a case there is no objection to the use of five letters as marks. The objection consists in arriving at these letters through the use of percentages which imply that the basic marking is done on the scale of one hundred. To arrive at the letter marks directly, assuming relative achievement, is of course a different situation and probably more defensible.

GRADING BY THE NORMAL CURVE

The normal curve in marking. The second general type of marking system employs the idea of the normal curve or probability integral. This method is sometimes known as the "Missouri Plan" because it was developed by Professor Max Meyer of the University of Missouri. In contrast with the scale of one hundred, the normal curve idea implies marking upon a basis of *relative*, not absolute, achievement. The underlying logic is that of chance or probability. It has been noted empirically that chance phenomena like the repeated tossing of a number of pennies tend to form a symmetrical bell-shaped curve something like that shown in Figure 13. The essential characteristics of such a curve are:

1. Most of the cases center rather closely about some central value or average.

2. Large deviations from such an average occur less often than small deviations.

3. Deviations of a given magnitude are equally likely in either direction from the central value.

4. The expectancy is unlimited at either extreme. In other words, the curve never, in theory, quite reaches the base-line but extends indefinitely in both directions. (The mathematician calls such curves asymptotes.)

Such curves summarize rather well many phenomena of biological variation. Since the distribution of mental abilities may quite logically be regarded as biological phenomena in their essential nature, it is not at all surprising that the study of individual differences to date has yielded many distributions which are sufficiently normal to justify statistical treatment upon the assumption of normality. At the same time it must never be overlooked that the normal distribution *need not fit* such data at all closely. To speak of the distributions of a random group of pupils with respect to general intelligence or ability in geography as "normal," really implies nothing more than a conviction based upon experience that such abilities appear to be distributed in much the same manner as we commonly observe to be the case with chance phenomena. Very few, if any, mental traits have been found to be distributed in a manner markedly at variance with the normal cure.

To return to the four characteristics of the normal curve as listed above, we find that they, individually and collectively, summarize pretty well our own experience and observation on the distributions of ability as found for the various school subjects. At any rate, there is little reason to believe that the assignment of marks upon a basis of the normal curve will introduce errors which are large compared with known errors in marking by other common schemes.

The real difference, for present purposes, between the 100-point scale and the normal curve rests in the fact that

the former assumes absolute standards of judgment, and the latter attempts nothing more than relative judgments. The former attempts the precision of physical measurements with their exactly defined units such as inches, hours, meters, pounds, etc. The latter attempts merely to arrange pupils in general order of merit; i.e., it is essentially a ranking process.

Under plans based upon the normal curve there is no attempt made to brand a pupil as "80" (or 80%) or as "87" (or 87%) with 100 as a point of reference. Instead, pupils are marked A, B, C, . . . or 1, 2, 3 . . . to four, five, six, or seven (occasionally more) letter or numerical marks.

Marking schemes are essentially matters of definition. It is necessary at this time to interrupt the discussion of marking schemes momentarily to bring home certain facts. After all, any marking scheme is arbitrary. Without definition it can have no meaning. If a school system is to develop a defensible plan for evaluating pupils' accomplishment, two things are absolutely essential, viz.,

1. The pupils must be placed in correct, relative positions or ranks with respect to each other; and,

2. The adopted marking scheme must be defined. Its sole meaning and value rest upon its definition to pupils, teachers, and parents alike.

With respect to the first of these requirements we need add little to the discussions of preceding chapters. This entire volume has held constantly in the foreground the need for valid and reliable measurement. If an examination, a series of examinations, or the composite evidence upon which marks are determined rank the pupils in the approximately correct order of achievement, it matters little what form the final statement of marks may take. The rest of the question is largely a matter for local decision. No marking system can be any more valid than is the underlying ranking of the

pupils. It is doubtful whether educational measurement thus far has risen much above the plane of rank-orders. If it has, the standard test alone reaches such a refinement. If the teacher has an adequate basis for evaluating her pupils in rank-order, almost any marking system, if defined, is a good one. If the basic data for marking pupils are invalid or unreliable, no marking plan will introduce the slightest refinement into these basic determinations.

This leads us to our second point, viz., that, granted approximately correct relative evaluations, the exact marking system reduces to a pure matter of definition of terms. An "A" cannot possibly have any other meaning than something analogous to "that level of achievement which, in a large and unselected group of pupils, is attained by the *best* five (or some other arbitrarily chosen) per cent of pupils." Indeed, until we have some exact unit of educational measure (equivalent to the physicist's foot-pound or watt-hour, etc.), what other meaning can a mark have? It is the author's contention that any teacher who can say, with proof, that Henry Jones is the best pupil in the class, Mary Ellen Brown stands second, Marion Smith is third, and so on down to Franklin White who stands last, need not concern herself greatly whether her pupils are marked with letters, numbers, words, or what. The exact and final recording of the facts at hand will be nothing more or less than a definition of practices. The normal-curve idea, the percentage plan, or any other scheme is fundamentally no better nor no worse than the relative rankings possible by means of the evidence at hand. For the present, educational measurement is rank measurement, not measurement in the physical sense. Almost any scheme of recording marks, provided it be adhered to by all teachers in the same school or school system, and provided further that it be understood by all concerned, will prove adequate if there is a valid and reliable provision for the measurement of the relative abili-

ties of the pupils to be graded. At the same time, the definition of local practices is essential in order that there be meaning to the final marks given.

It is unfortunate that diversity exists in marking systems from city to city and from state to state. A uniform plan would have many advantages. On the other hand, a uniform marking plan throughout the United States would have certain inaccuracies, since we have reason to believe that there are variations in the quality of school work from one locality to another.

Normal distributions of school marks. Assuming that the better marking systems are those employing relative rather than absolute standards, of what use can the normal curve be? In the first place we should disabuse ourselves of the idea that the normal curve (or any other mathematical concept) will tell us exactly *how many* pupils should receive A, B, C, etc. The division of a group of pupils among the several letter or number marks is largely a matter of definition, but not wholly so, as we shall see. If we assume that normality of distribution is sufficiently true to the facts of human nature, the normal curve does lead us toward certain decisions. The fact that observed distributions of mental abilities are ordinarily bell-shaped, bunched in the middle, and tailed out gradually toward the extremes does furnish a rough guide to the relative assignments of different marks. Thus, to adopt a grading system like the following would hardly be defensible in the light of all known facts.

Letter mark	A	B	C	D	E
Per cents obtaining	20	25	40	10	5

Such a plan runs counter-current to the probable symmetry of the real abilities of a large group of pupils. A visibly better plan would be:

Letter mark	A	B	C	D	E
Per cents obtaining	5	25	40	25	5

We are left with a rather difficult problem of deciding exactly how the various letter or numerical marks should be distributed. In the absence of any better basis, students of this question have gradually come to think that the proportions may well follow those to be expected in chance phenomena. If we can defend such a position, there is a simple mathematical approach to the solution of our problem, *viz.*, through the binomial theorem.

If we toss four pennies for a large number of times, we discover that all possible outcomes fall into five categories. Moreover, each of the five categories tends to show roughly fixed proportions of relative occurrences. We speak of these relative proportions as the expectancies or probabilities.

Calling H heads and T tails, by expanding $(H+T)^4$ we get:

$$H^4+4H^3T+6H^2T^2+4HT^3+T^4.$$

The coefficients of the five successive terms (categories) are 1, 4, 6, 4, and 1, summing to 16. Expressing the coefficient of each term as a fraction of 16, we can build up a table like the following:

Category	Probability of Occurrence			
Four heads—no tails	1 in 16	or	about	6%
Three heads—one tail	4 in 16	or	about	25%
Two heads—two tails	6 in 16	or	about	38%
One head—three tails	4 in 16	or	about	25%
No heads—four tails	1 in 16	or	about	6%

If we should plot these expected results as a graph, it would be noted that the results run more or less parallel to the characteristics of the normal curve, i.e., a heaped symmetrical distribution which tails out gradually toward either end. Such a distribution is not normal for a number of reasons. However, if the interested student will expand the binomial $(H+T)^n$ for successive values of n, plotting each expansion as a curve, he will probably be convinced that the *limit* of the binomial, when n becomes infinite, is the curve

which we have discussed under the name of the normal curve or probability integral.

Below are given the expansions of the expression $(H+T)^n$ for values of n between 2 and 6. Below each term is given the coefficient of that term expressed as a fraction of the sum of the coefficients of all terms.

$$(H+T)^2 = H^2+2HT+T^2$$
$$\frac{1}{4} \quad \frac{2}{4} \quad \frac{1}{4}$$

$$(H+T)^3 = H^3+3H^2T+3HT^2+T^3$$
$$\frac{1}{8} \quad \frac{3}{8} \quad \frac{3}{8} \quad \frac{1}{8}$$

$$(H+T)^4 = H^4+4H^3T+6H^2T^2+4HT^3+T^4$$
$$\frac{1}{16} \quad \frac{4}{16} \quad \frac{6}{16} \quad \frac{4}{16} \quad \frac{1}{16}$$

$$(H+T)^5 = H^5+5H^4T+10H^3T^2+10H^2T^3+5HT^4+T^5$$
$$\frac{1}{32} \quad \frac{5}{32} \quad \frac{10}{32} \quad \frac{10}{32} \quad \frac{5}{32} \quad \frac{1}{32}$$

$$(H+T)^6 = H^6+6H^5T+15H^4T^2+20H^3T^3+15H^2T^4+6HT^5+T^6$$
$$\frac{1}{64} \quad \frac{6}{64} \quad \frac{15}{64} \quad \frac{20}{64} \quad \frac{15}{64} \quad \frac{6}{64} \quad \frac{1}{64}$$

If we collect these results as a table, changing the stated fractions to per cents and denoting the categories by the letters A, B, C, etc., we have the following as the approximate expectancies of the occurrences of each letter-grade upon the assumption of chance distribution of pupil abilities:

No. of Letter Grades Employed	Letter Grades						
	A	B	C	D	E	F	G
3	25%	50%	25%				
4	12%	38%	38%	12%			
5	6%	25%	38%	25%	6%		
6	3%	16%	31%	31%	16%	3%	
7	2%	9%	23%	31%	23%	9%	2%

The per cents in the foregoing tabulation have been rounded off in such a fashion as to keep the sum in each series one hundred per cent. Note that if n equals the number of letter-grades to be employed, the expectancies are found from expanding the binomial to $n-1$ as the exponent.

That such distributions are roughly the normal expectancies as well can be determined by consulting any table of the normal curve or probability integral. Any textbook on statistical methods gives such tables.

Limitations of normal distributions of marks with small classes. Few persons will find any serious quarrel with the above distributions of marks as a matter of pure theory, since we have attempted nothing more than a definition of terms upon the assumption that pupils distribute themselves according to chance, and that, if this be the case, theorems of probability furnish rough guides to a defensible marking plan. The derivation of these relative proportions is based upon one very important additional assumption, viz., *that n is very large.* This means, in effect, that such distributions of markings may be held to be fairly valid for large numbers of pupils, but that they must not be applied too literally to small groups. In general, it is unwise to apply such systems in a purely mechanical fashion if fewer than one hundred pupils are to be marked as a unit. Even with a few hundreds of cases, some injustice is likely at times, although such errors are probably very slight in comparison with numerous other non-eliminable and ever-present errors in grading such as subjectivity, unreliability of examinations, etc.

It is generally known that the average accomplishment of typical classes of ten to fifty pupils vary somewhat from year to year. Occasionally very large variations will occur, although these are the exceptions, not the rule. Teachers generally are prone to over-estimate the amount of variation to be expected in the average abilities of successive classes. To illustrate the probable amounts of variation from one class to the next, we may assume certain facts:

1. That each class contains thirty-six pupils.

2. That the same objective and rather reliable test was given to each successive class.

3. That the average score of the first class on the test mentioned was 90 out of a possible 150 points.

4. That the standard deviation of the marks was approximately 30, and the probable error was approximately 20. (See Part IV for the meaning of these terms.) For present purposes the meaning of the probable error will be sufficiently defined if we assume that within 20 points on either side of the average (90), i.e., between 70 and 110, roughly fifty per cent of the scores fall.

Our problem is to decide within what limits the averages of successive classes will probably fall. To do this we need what is called the probable error of the average. The usual formula is:

$$PE_{(Average)} = \frac{PE_{(Distribution)}}{\sqrt{N}}$$

Substituting the assumed values, we have:

$$PE_{(Average)} = \frac{20}{\sqrt{36}} = \frac{20}{6} = 3.3$$

Upon the basis of our calculation we may conclude that the chances are 50:50 that next year's class (of thirty-six pupils) will show an average score on the same test within the limits of plus or minus 1 PE, i.e., between 86.7 (90−3.3) and 93.3 (90+3.3). Without troubling to explain the mathematics of the situation, we can say:

The chances are 50 in 100 that the average of any succeeding class will fall between 86.7 and 93.3

The chances are 82 in 100 that the average of any succeeding class will fall between 83.4 and 96.6

The chances are 98 in 100 that the average of any succeeding class will fall between 80.1 and 99.9.

There is thus but about one chance in fifty that the average of a succeeding class will depart by as much as ten points in either direction from the average of the first class. It may

seem that ten points is a large departure. Just how large it is may now be estimated. We started with the assumption that the middle fifty per cent of the class scored between 70 and 110, a range of forty points. In such a case the total range of scores of the individual pupils of the class would hardly be less than 100 points. The average from class to class would fall roughly ninety-eight times out of one hundred between 80 and 100, or within a range of twenty points, which is about one-fifth of the difference between the best and poorest pupil in the class. About four times out of five (82 in 100) the average, from class to class, would fall between 83 and 97 (83.4 and 96.6 to be more exact), a range of fourteen points, or about one-seventh of the difference between the highest and lowest score in a given class.

Without trying to minimize the seriousness of such fluctuations in averages, it is necessary to call attention to the fact that, after all, when the final marks of the class are distributed upon a five- or seven-letter grade basis, it is easy to exaggerate the seriousness of variations from class to class. We should attempt some method of making allowance for differences in the average achievements of small classes. This question will be the next issue to be discussed here.

Proposals for allowing for variations arising from small samplings of pupils. A given class of twenty or thirty pupils may well be regarded as a small sampling of the pupils of a number of successive classes. Many teachers have objected violently to the use of the normal curve in grading small classes, upon the argument that such marking is too mechanical and arbitrary. Such teachers wish, with much justification, to use their judgment in departing from the adopted proportions of the letter-grades agreed upon as the basis of marking in that particular school. With such a desire the author is in considerable accord. Nevertheless there are a number of considerations which must be kept in mind.

1. In the absence of *demonstrated* departures of a given class from typical conditions, the teacher should be very cautious about departing far from the adopted scheme. Human judgments are so fallible that small variations cannot be detected unless rather exact and objective measurements are employed. Large variations will ordinarily be open to casual inspection; small variations will not. It is probably a safe rule to adhere to the adopted plan unless the judgment of the teacher is supported by facts obtained from the use of well-constructed tests, standardized measures, or intelligence testing, etc.

2. While variations in classes occur from year to year, it is unlikely that these occur constantly in the *same* direction. Over a period of years the pooled marks should approximate fairly closely the adopted scheme, even when certain classes are allowed to depart noticeably from the plan in use. More times than not, teachers fall into the habit of "always having superior (or, less often, inferior) classes"; i.e., each year they diverge in the *same* direction from the adopted plan. This phenomenon may be a subtle form of self-flattery in many instances. The teacher who keeps no cumulative records will nine times out of ten either (*a*) use the grading system in an arbitrary and mechanical fashion, or (*b*) show systematic and continual departures from the adopted scheme. Neither of these outcomes can be defended.

3. The dangers from a too-rigid adherence to any grading plan with small classes are undoubtedly real; at the same time variations in classes may be rather insignificant in the light of other sources of error in marking. In particular it is to be noted that no grading plan will make matters markedly better or worse than the inaccuracies present in the bases by which the marks are decided, whether these be recitations, written work, examinations, standard tests, or what. If there is an adequate basis for marking pupils so that they are arranged fairly reliably in order of achievement, almost any grading system will work reasonably well.

4. The fundamental issue is to provide objective, valid, and reliable instruments for determining relative ability. This is essentially a matter of ranking pupils in order of relative merit. Secondarily there is the problem of changing these rankings into some literal or numerical system of marks. This second step cannot introduce any refinement into the first and more fundamental consideration. Then we can attempt, more or less successfully, the third problem of allowing for the effects of small classes.

5. Any grading plan is largely a matter of definition. It is a "gentlemen's agreement" to play the game by the same rules. A teacher who, year after year, departs systematically from this agreement is a "poor sport"; she may be justly accused of willfully refusing to co-operate in defining marks in the minds of pupils and parents. To excuse the practice on such grounds as having "high" standards (these teachers never apparently have "low" standards), or that she gets better results from her pupils than do other teachers in the same school, only puts the teacher in a bad light, no matter what the facts may be.

We may now consider briefly certain proposals for allowing for variations in average abilities of small classes from one year to the next or from one subject to the next.

First Proposal. In stating the approximate percentages to be given each letter grade, allow some latitude for the teacher's final judgment. Thus:

A	B	C	D	E
5%–10%	20%–30%	35%–45%	20%–30%	5%–10%

Although some latitude is essential in order to avoid slavish adherence to the scheme, there is always the danger that teachers will abuse the privilege and depart systematically in one direction or the other. There is the further objection that the mere allowance of latitude does *nothing at all* toward providing some defensible basis of setting the scale to fit particular classes.

Second Proposal. Check the direction and amount of variation or selection present through the administration of an intelligence test. This is, in part, the recent proposal of Ellis,[1] although many earlier advocates of this idea are to be found. Although this method has undoubted value, it is open to several objections: (*a*) the extra cost, (*b*) the extra labor, and (*c*) most important, intelligence is often a poor index to success in certain school subjects, not to mention the possibility of classes working well above or below the indications of their intelligence ratings. It is unlikely that teachers generally will be convinced that Ellis's plan is the best solution in sight, although this plan has its merits.

Third Proposal. Administer a good standard test as a check on your assignment of letter grades. Symonds[2] favors this plan, and he has set up the needed machinery for handling the method with a number of high-school tests.[3] Symonds's plan is probably much to be preferred over that of Ellis, as it represents a direct attack on the problem, and he has worked out convenient tables for handling many high-school test results. No such data are available for elementary-school subjects. Moreover, it is only fair to state that Symonds's work is not directly aimed at our present proposal since he is primarily concerned with a method of transmuting standard test scores into grading units. His tables do furnish valuable checks upon such questions as the variation and selection present in a small class.

Fourth Proposal. The author has reached a somewhat different solution of the question of allowing for variation and selection in marking successive classes or different sections of the same class. The real problem is that of throwing

---

[1] R. S. Ellis, *Standardizing Teachers' Examinations and the Distribution of Class Marks* (Bloomington, Illinois: Public School Publishing Company, 1927), pp. 112-114.

[2] P. M. Symonds, *Measurement in Secondary Education* (New York: The Macmillan Company, 1927), pp. 507-529.

[3] P. M. Symonds, *Ability Standards for Standardized Achievement Tests in the High School* (New York: Teachers College, Columbia University, 1927).

successive (and possibly varying) classes or sections upon a single scale so that the direction and amount of departure of any class or section from the central tendency of many such groups may be noted. This, in the author's opinion, may be done with a degree of accuracy commensurate with the needs of the situation through the construction and use of duplicate or equivalent examinations. The building of two or more forms of the same examination so as to capitalize on the equalizing force of chance was described in Part II of this volume. In brief, the plan is to prepare hundreds of items and then to deal them by chance into from two to five piles, depending upon the number of duplicate forms adjudged necessary. If each resulting form contains at least one hundred items, the averages of such chance forms will ordinarily differ by not more than from two to five score points, an amount of inequality of no great practical consequence in view of the other unavoidable errors present in any marking scheme.

If there are several sections of the same class, different tests or examinations may be given to each, and yet the resulting scores may be thrown into a single distribution for purposes of determining marks. Even if the classes are sectioned upon a basis of ability, the duplicate examinations may be planned so that the pupils from all sections are placed upon a single scale of measurement. More will be said later on this point.

When there is but a single section a semester or a year, the present proposal offers no assistance at the outset. Over a period of two or three years the effect is to generate a set of local norms. As time goes on, these norms become more and more accurate if cumulative records are kept. The proposals of Ellis and Symonds, particularly the latter, may at the outset be used to advantage, but will gradually become unnecessary. Our proposal reduces to the very simple proposition of avoiding the pitfalls inherent in the grading

of small classes by pooling successive sections or classes until the resulting numbers are large enough to be viewed as stable samples. The principal advantage of this proposal is that it is cheaper and less laborious than the plans advocated by Ellis or Symonds.

Before closing the section on the values of duplicate forms, we must anticipate a possible misunderstanding. The reader may say that, after all, his or her regular practice of building a new test each semester is exactly equivalent to the present proposals. Not at all! It is to be hoped that the author has been sufficiently lucid to dispell such a misunderstanding. The construction of two, three, or more forms of an examination *at one time* is essentially different from the production of the same number of "forms" at *different* times. Simultaneous duplication capitalizes on the equalizing force of chance (or chance plus judgment); successive duplication with anything like approximate equality is an almost impossible task unless extended experimentation is resorted to.

The problem of allowing for the variation and selection inherent in small classes is aggravated by the practice of sectioning classes upon a basis of ability. The next section comments upon the marking of such sections.

The marking of classes sectioned upon a basis of ability. It is now common practice to divide large groups of pupils into ability sections. These are often designated as "X," "Y," and "Z" groups. The basis for such sectioning is most often the scores from mental or educational tests. When sectioning upon a basis of ability is used, the sections are likely to present the situation shown in Figure 14.

Figure 14 shows two significant facts: (*a*) that the average abilities of the three groups are quite different, and (*b*) that there is much overlapping of the abilities of the three groups. In view of these facts, we must make some allowances in our grading scheme for such differences in ability.

Practices are sharply divided upon the issue of marking superior, normal, and sub-average sections of the same subject. Two general plans are used.

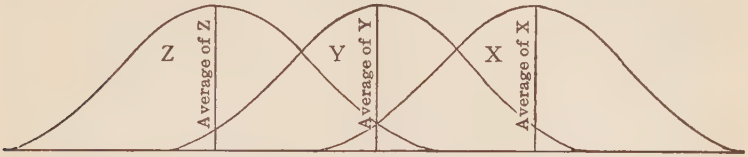


FIG. 14.—Showing the effects of sectioning a group of pupils into three ability groups.

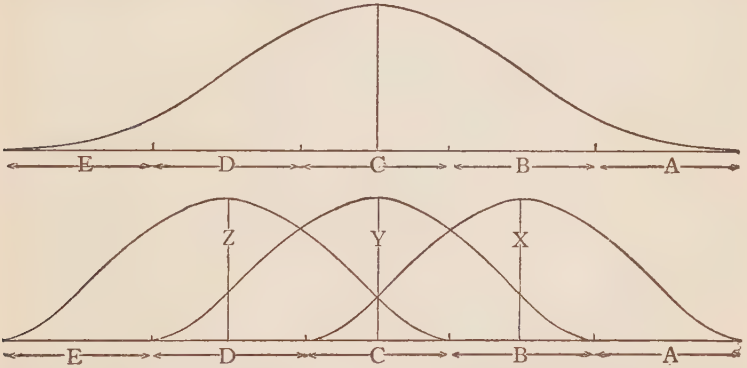


FIG. 15.—Showing the effect of marking classes sectioned upon a basis of ability when a single standard of marks is employed. The upper distribution shows the marks of the total group (the three sections pooled) and the lower distribution shows the marks for the X, Y, and Z sections separately.

Plan I. Mark all sections upon a *single* basis. This means that the X (superior) section will receive a larger proportion of high marks than will the Y (average or normal) section, and, in turn, the average section will receive a larger number of high marks than will the Z (sub-average) section. Figure 15 shows roughly the effect of this plan. The upper distribution shows the marks which might be given to the total group upon a basis of five letter-grades. The lower

distribution shows how the marks might be apportioned after the total group has been broken into three sub-groups on a basis of ability. In drawing these curves it is assumed that the sectioning reduces the abilities of each section to a range of three letter-grades (in contrast with a range of five letter-marks for the total group).

Plan II. Mark each section on a *different* basis; i.e., make the marks for each section show merely relative accomplishment *within* that section, thus ignoring the fact that the three sections represent different levels of ability. In this case all sections would receive equal numbers of each letter-grade. Some schools follow such a plan, but designate the letter-grades with subscripts to indicate the section in which the mark was earned thus: A_x, A_y, A_z, or B_x, B_y, B_z, etc. It is obvious that an A or a B does not have the same meaning in a Z section as it does in an X section. There are just as many standards of marking under Plan II as there are ability groups or sections. It should be noted that this plan requires (logically) that there be as many failures among the brightest pupils as there are among the dullest ones.

The choice between Plans I and II is an arbitrary one. There is no doubt that the first plan is superior from the standpoint of the theory of measurement. It represents measurement by a single, fixed, and defined standard in something of the sense of measurement in the physical sciences. On the other hand, Plan II offers important educational advantages in that it is less likely to arouse the conviction (in Y and Z sections) that high marks are almost, if not quite, impossible. Many teachers feel that a pupil doing good work, relatively speaking, in a slow (Z) section should be allowed to earn an A or a B exactly as if he were a member of a fast (X) group. There is something to be said in favor of such a view from the standpoint of motivation, but there are certain weaknesses in this plan that should be considered. In the first place, the more able students some-

times become aware of the facts and object to a situation which makes them do, perhaps, twice as much work to earn an *A* as is the case with a pupil in a Z section. In some instances students have been known to "lie down" in taking the tests used for sectioning in order that they may earn a berth in a slower-moving section. Again, schools need at times to make further evaluations of pupils, e.g., in recommending high-school graduates to college registrars or to prospective employers. Thus far many schools employing Plan II have not yet shown the courage of their convictions by granting differentiated diplomas or certificates of completion. In the State of California only A and B grades in high school are "recommending grades" to first-class colleges and universities. Such A's and B's, of course, cannot be earned in Y or Z sections. Under the California plan it is inevitable that high-school pupils will become aware of the significance of ability grouping. Viewed from any angle it seems impossible to employ Plan II without ultimately "laying the cards on the table" so that an A or a B may be construed by all concerned as having meaning only with reference to the section (ability grouping) in which it was earned. Pupil, parent, teacher, and employer will have to know the facts. When this condition exists, Plan II will lose much of the advantage claimed for it by way of motivation.

Plan II suffers in comparison with Plan I in that it requires separate examinations for each section. Under Plan I uniform examinations may be set for all three (or other number of) sections. In fact, uniform examinations for all ability groups are essential if all pupils are to be marked upon a single scale. Such tests or examinations must be constructed so as to include the basic contents of the instruction in all three sections. Examinations of this nature are quite practicable. It is only necessary to include, as the first few items of such an examination, questions which will be passed

by all slow or Z section pupils, and then to proceed by degrees to increase the difficulties up to a final point where the items are failed by most or all of the fast or X pupils. Some teachers will object to this idea without realizing that this procedure is exactly what occurs in the case of any standardized mental or educational test. It is true that the Z sections will be exposed to test materials which have never been taught. This is not a fundamental objection in comparison with the obvious advantage of having all sections evaluated upon a single scale of accomplishment. In no other way is it possible to throw X, Y, and Z groups into direct comparison, one against the other.

MEASUREMENT AND RANKING

Measurement as rankings. *Before any norm need be consulted, before any interpretation need be sought, and before any sort of evaluation is justified, we must first establish the fact that the test employed is capable of arranging the pupils of a class in a valid and reliable rank-order of ability.*

Any test, no matter how worthless, yields a series of scores for the members of a group of individuals. The fact that a wide range of scores is obtained is quite without meaning unless certain facts are first established. Very pretty sets of scores for a class of pupils may be obtained by "shooting" dice or tossing pennies. Giving a test is something like a horse race. Some horse always wins, and the rest come in second, third, fourth, etc., until the last horse crosses under the wire. The question is: "Did the best horse win?"

A good test arranges the pupils in the approximate rank-order of ability. There is little error in the ranking. *A worthless test (zero reliability) arranges them in an order no better nor no worse than a lottery which is "on the level."*

We thus have a picture of the two extremes of the scale of reliability, a perfectly reliable test and a worthless or entirely unreliable one. The former (if such existed) would

yield a set of scores for a class such as would justify each and every score being taken at face value. If the highest score were 89, the next 86, the next 77, and so on, we could say without danger of error that the pupil obtaining 89 was best, the 86-pupil was next in ability, the 77-pupil was third, and so on; all this assuming that the test was perfectly reliable. Suppose on the contrary that the test was absolutely unreliable (zero in reliability). Any set of scores obtained would be no more accurate than having the pupils draw the same numbers from a hat.

In actual practice, tests are seldom near zero in reliability, and they usually are far from perfect in reliability. It follows that all test scores contain errors in greater or less degree, or as test-workers say, the test is *fallible*. A fallible test is the only kind the teacher can ever hope to administer. It follows that every test score must be taken *cum grano salis*. If we arrange the pupils of a class in rank-order upon the basis of any fallible test, it is certain that some pupils will be interchanged in rank. The best that we can hope for is that the displacements will not be very large in any single case, and that the average displacement will be as small as possible.

Consider now a possible situation. It is near the end of a school term. The teacher is beginning to think about making a final examination, the results from which will enter importantly into the final grades. She decides upon a certain kind of test, old-type or new-type; it matters little. The questions are phrased, written down, and administered. The test is scored, and a wide range of scores is obtained. What do the scores mean?

Before attempting to suggest an answer to this question, let us propose an innovation to the examination practice for this year. Suppose the teacher should decide to build a second test equivalent to the first so far as she is able to judge. To verify the results of the first test, she gives the

second one which she has constructed. The results are scored and tabulated, pupil by pupil, alongside the first set of scores. They disagree! In places the differences are disturbingly large. At other times they are too slight for serious attention.

The teacher now calculates the averages on the two examinations. One shows an average of 80 and the other of 71, a difference of nine points.[1] Also, the range of scores in the first test was from 72 to 95, and in the second from 56 to 91. Thus, both the average and the variability (range) differed in the two instances.

What would such a result mean?

It is apparent that the numerical scores are seriously upset by the two facts already mentioned: the differences in averages and the differences in variability. Would the ranks be as markedly disturbed by such facts? Not necessarily so. In fact, a few minutes' thought will probably convince the reader that rank-orders are less subject to inequalities from one equally valid but unequally difficult examination to the next than are the absolute numerical scores. If this be true, there is an important lesson for us here in evaluating pupils' achievement.

An actual example. We can now turn our attention to an actual example. Two examinations were given. So far as inspection could show, they were equally difficult. One was given on Monday, the other on the following Wednesday. Each examination contained 100 objective items. Table 73 shows the results, together with certain computations.

Discussion of Table 73. The conditions of the experiment must be kept in mind as we study this table. In the first place, this experiment differed from the ordinary classroom

[1]This figure is chosen because it represents about the average difference found in a series of similar double-examinations studied experimentally by McGregor and Ruch. *See Objective Examination Methods in the Social Studies* (Chicago: Scott, Foresman and Company, 1926), especially pp. 6-12. Or, *see* Chapter III of the present volume for an abstract of McGregor's study.

situation only in the fact that *two* examinations were given instead of the usual *one*.

These two examinations were equally carefully constructed, and so far as the teacher could judge, they were equally valid and reliable. They were made at an interval of several days, and, of course, there was no way of making them

TABLE 73

SCORES AND RANKS OF A CLASS OF 20 PUPILS ON TWO OBJECTIVE TESTS GIVEN ON TWO DIFFERENT DAYS

PUPIL	SCORE ON TEST I	SCORE ON TEST II	DIFFERENCE	RANK ON TEST I	RANK ON TEST II	DIFF. IN RANKINGS
1	38	56	18	20	18	2
2	60	76	16	15	14	1
3	76	88	12	9	8	1
4	89	99	10	2	2	0
5	67	78	11	12	12	0
6	82	87	5	6	9	3
7	49	57	8	17	17	0
8	64	68	4	13	15	2
9	58	77	19	16	13	3
10	85	89	4	4	7	3
11	86	94	8	3	4	1
12	83	97	14	5	3	2
13	71	79	8	11	11	0
14	46	54	8	18	19	1
15	72	92	20	10	5	5
16	61	64	3	14	16	2
17	41	46	5	19	20	1
18	91	100	9	1	1	0
19	81	91	10	7	6	1
20	80	86	6	8	10	2
Averages	69.0	78.9	9.9			1.5

exactly equal in difficulty. The matter of equality of difficulty was again a question of the opinion of the teacher; she made them as nearly equal as she could judge.

The two examinations covered the same general scope of classwork and appeared to be interchangeable in use; i. e., so far as the teacher could tell, one was just as good as the other, and she was quite willing that either should have been

used *alone* as a measure of her pupils. This point has an important bearing on our discussion.

These two examinations were very reliable for tests of 100 items. The reliability coefficient was found to be 0.95.[1] The conditions of this experiment would be analogous to the situation of a test having been given but the test papers accidentally destroyed. A re-test was made up to be as nearly equivalent to the first as possible. The question is: What difference in the scores of individual pupils might be expected?

With these facts in mind, we can turn to the results shown in Table 73.

1. We note that Test I averaged about ten points harder (lower) than Test II, a common finding under such circumstances.

2. No pupil obtained exactly the same score on both tests. The least difference found was three points, the largest disagreement was twenty points, and the average disagreement was about ten points (9.9).

3. When the actual scores were changed into rank-orders, five pupils received exactly the same rank on both tests. The smallest difference in ranks was zero, the largest proved to be five, and the average was 1.5. The differences in ranks were much smaller than the differences in the actual scores.

4. The tests used in this experiment were unusually reliable (0.95) in the light of actual practice. There can be no exaggeration of the facts on this score. The one possible element in the situation which may not be typical is the fact that one test averaged ten points easier than the other. Extensive investigations have shown that average differences of from eight to twelve points are roughly the expectancy when dual examinations are given.[2]

---

[1]See Chapter XV for the meaning and calculation of reliability coefficients.
[2]See Chapter III.

5. When two independent (but supposedly equivalent) examinations are given, one or all of three things may happen:

(*a*) The tests may prove to be of unequal difficulties (i. e., yield different average scores, as in the present case).

(*b*) The tests may show unequal variabilities (i. e., the range or spread of score may be quite different).

(*c*) There may be little correlation or correspondence on the scores of individual pupils (i. e., the scores might show very different rank-orders on the two tests).

6. In the present case, (*a*) was found to be true, but inspection will show that (*b*) and (*c*) are not serious disturbing factors. All in all, therefore, we have a situation which can hardly be held to be out of the expectancy.

Ranks vs. per cents. The test under study was purposely made to yield a maximum of 100 points. This was done in order (1) to make it comparable with the ordinary examination which is usually graded upon a basis of 100 (or 100%), and (2) to raise an issue about the 100-point (100 per cent) grading plan.

The previous discussion in this chapter has already pointed out that there are two general types of grading practices employed widely at present: grading from 0 to 100 (often stated as per cents) and grading by some ranking system based upon the normal curve, e. g., stated percentages of A, B, C, etc., or some equivalent designation.

It is also quite evident from the previous content of this chapter that the merits of these two rival methods have been the subject of voluminous controversy. The present experiment contributes data pertinent to the issue.

Numerical (obtained) test scores or examination marks are used in two ways: (*a*) as they stand (63, 78, 91, etc., or sometimes 63%, 78%, 91%, etc.), or (*b*) translated into

letter grades by some such plan as the one which is outlined below.

A	95 or above	PLAN A
B	85 to 94	
C	75 to 84	
D	70 to 74	
E	(Failure) Below 70	

The exact numerical points of division naturally differ somewhat from teacher to teacher or from school to school. For our purposes, it matters little where the breaks come.

The second (and newer) method of assigning grades is based upon the idea of the normal curve (probability curve). *It is essentially a counting-in process.* Stated percentages of each letter-grade are given, regardless of the exact magnitudes of the original obtained numerical grades. This method is fundamentally a ranking method. A common scheme (and the one adopted for comparison here) is:

A	Highest 5%	PLAN B
B	Next 20%	
C	Middle 50%	
D	Next 20%	
E	(Failure) Lowest 5%	

For convenience we have termed these two plans, Plan A and Plan B. Tables 74 and 75 show the two plans applied to the data of Table 73.

Discussion of Tables 74 and 75. It will be necessary for us to keep our bearings in discussing Tables 74 and 75. In the first case we must exclude the question of whether the counting-in (normal curve) idea is strictly applicable in comparing *different classes* of pupils. This point has already been discussed in a previous section. We are concerned here with *successive (and supposedly equally valid) examinations of the same pupils.*[1]

[1]The problem here also is not equivalent to whether raw scores or ranks show the same correlation, or whether rank correlations are better or worse than correlations of actual scores. The two correlations are substantially the same.

TABLE 74*

PLAN A APPLIED TO THE TWO SETS OF EXAMINATION MARKS OF TABLE 73

LETTER GRADE	BASIS OF ASSIGNMENT	TEST I	TEST II
A	95 or above	None	No. **18** (100) No. **4** (99) No. **12** (97)
B	85 to 94	No. 18 (91) No. 4 (89) No. 11 (86) No. 10 (85)	No. **11** (94) No. **15** (92) No. **19** (91) No. 10 (89) No. **3** (88) No. **6** (87) No. **20** (86)
C	75 to 84	No. 12 (83) No. 6 (82) No. 19 (81) No. 20 (80) No. 3 (76)	No. **13** (79) No. **5** (78) No. **9** (77) No. **2** (76)
D	70 to 74	No. 15 (72) No. 13 (71)	None
E	Below 70	No. 5 (67) No. 8 (64) No. 16 (61) No. 2 (60) No. 9 (58) No. 7 (49) No. 14 (46) No. 17 (41) No. 1 (38)	No. 8 (68) No. 16 (64) No. 7 (57) No. 1 (56) No. 14 (54) No. 17 (46)

*Individual pupils are designated No. 1, No. 2, etc., to correspond with Table 73. The numbers in parentheses are the actual examination marks. The bold-face entries represent disagreements between letter marks by Tests I and II.

The question is that of illustrating whether actual numerical scores on the basis of 100 (or 100%) are as accurate as the method of counting-in (the normal curve idea) *when examinations, if repeated, show* (*a*) *differences in average difficulty, and* (*b*) *differences in variability, in addition to* (*c*) *those differences due to unreliability, proper.* Factor (*c*) has been controlled to a degree which make the present data far more reliable than would be true in the average run of events.

TABLE 75*

PLAN B APPLIED TO THE TWO SETS OF EXAMINATION MARKS OF TABLE 73

LETTER GRADE	BASIS OF ASSIGNMENT	TEST I	TEST II
A	Highest 5%	No. 18 (91)	No. 18 (100)
B	Next highest 20%	No. 4 (89) No. 11 (86) No. 10 (85) No. 12 (83)	No. 4 (99) No. 12 (97) No. 11 (94) No. **15** (92)
C	Middle 50%	No. 6 (82) No. 19 (81) No. 20 (80) No. 3 (76) No. 15 (72) No. 13 (71) No. 5 (67) No. 8 (64) No. 16 (61) No. 2 (60)	No. 19 (91) No. **10** (89) No. 3 (88) No. 6 (87) No. 20 (86) No. 13 (79) No. 5 (78) No. **9** (77) No. 2 (76) No. 8 (68)
D	Next 20%	No. 9 (58) No. 7 (49) No. 14 (46) No. 17 (41)	No. **16** (64) No. 7 (57) No. **1** (56) No. 14 (54)
E	Lowest 5%	No. 1 (38)	No. **17** (46)

*Individual pupils are designated as No. 1, No. 2, etc., to correspond with Table 73. The numbers in parentheses are the actual examination marks. The bold-face entries represent disagreements between letter marks by Tests I and II.

Moreover, the differences termed (*a*) and (*b*) are not extreme, as experiments have shown.

Plan A shows twelve disagreements in final letter grades for the two tests. Plan B shows but six (the cases in bold-face type). Plan B, of necessity, shows no disagreements in the *numbers* of each letter-grade given. Plan A shows many such variations in letter-grades from one test to the next:

DISTRIBUTION OF LETTER-GRADES BY PLAN A

	A's	B's	C's	D's	E's (Failures)
Test I	0	4	5	2	9
Test II	3	7	4	0	6

Let us assume that we have no other basis for assigning marks to these twenty pupils than the results of Test I and II. Which set is better? There is no answer. They correlate well, and we must assume that they are approximately equally good. Which plan for marking is the better? The answer seems to favor Plan B, both on the basis of the number of disagreements and on the basis of uniformity of different letter-grades given.

If the results of this experiment are typical, differences of difficulty (and of variability, if present) from one examination to another very largely destroy any validity which *fixed or absolute* standards seem to have in theory. (By fixed standards we mean stated passing marks, stated numerical limits of A's, B's, etc., and other direct uses of obtained numerical scores when these are turned into the 100%-scale of markings.) Note, however, that these objections to the actual numerical scores are not based upon their inherent value, reliability, or validity, but upon the attempt to *transmute* such actual numerical scores into *per cents* (or in general the familiar scale of marks beginning at 0 and extending to 100).

A perfectly valid and reliable test would rank every pupil in a group in exact order of merit. Moreover, if the test were perfectly valid and reliable, a pupil obtaining 90 would probably be more superior to one obtaining 85 than the latter would be to one obtaining 84, but these scores would not represent per cents for many reasons, among which are:

(1) The zero point of such a series of scores is unknown.

(2) Such a scale ends at 100 only if the maximum score *happens* to be 100. Had there been 113 items, 113 would have been a perfect score. Had there been 77 items, no pupil could have earned "100."

(3) Had a harder test been given, the scores would have averaged much lower. To answer all the questions on an easy test is to accomplish less than to answer ninety out of a hundred on a very difficult one.

(4) The units of scores on tests and examinations are arbitrary. Where they begin and end depends upon the difficulty of the test. Moreover, the units change in value from one part of the test to the next, the first ten points are usually easier to obtain than the last ten points. Test scores are not numbers in the arithmetic sense which permits addition, subtraction, multiplication, division, and other operations regardless of whether the numbers are near zero or are very large.

(5) Unless much is known about the meaning of the numerical scores on a test (and seldom is this condition well met in practice), it seems to be safer to regard the scores as rank-orders.

(6) The more reliable the test, the more the resulting ranks may be taken at face value.

(7) The 100%-scale should probably be abandoned in favor of methods more nearly approximate to the known facts about individual differences. This is especially true if objective tests are employed where the numerical scores are functions of such facts as the length of the test, the difficulty of the test, etc.

Part IV

STATISTICAL TREATMENT AND INTERPRETATION OF OBJECTIVE TEST RESULTS

CHAPTER XV

STATISTICAL PROBLEMS RELATED TO MEASUREMENT[1]

Introduction. A number of statistical measures like the standard deviation and the coefficient of correlation have been mentioned freely in this volume, particularly in Part III. Some of these have received casual and superficial definition in passing. This chapter brings together as seven *Problems* the basic statistical processes and concepts necessary to a reasonably secure mastery of the principles of test interpretation.

Although the time has not yet come when the writer on educational measurement dares to take for granted that all teachers are thoroughly conversant with elementary statistical concepts, we have certainly reached the stage where no apologies are necessary for introducing such topics.

The discussions of this chapter center about a very few elementary procedures in the application of statistical method to such problems as the summarizing of test scores, the critical evaluation of tests, the determination of reliability and the accuracy of individual test scores, and the interpretation of the significance of correlation coefficients as measures of relationship and prediction. It is to be hoped that the reader will consult the references mentioned in the footnotes and the *General Bibliography*.

PROBLEM 1

Summarizing a Series of Test Scores

All teachers are already familiar with the arithmetic mean or "average" and with its general uses and significance.

[1]This chapter has been purposely reserved to the end of the volume so that it may be omitted without much loss to the general reader. The student will probably find it very much worth while to read through the discussion of the seven *Problems* in this chapter.

Most educators have also come to think easily in terms of the median or mid-measure as well. The arithmetic mean (ordinarily called the average or simply the "mean") and the median are the two most common *measures of central tendency*. The expression "central tendency" is worth knowing since its obvious meaning helps to define the concept of the average or the median.

Ordinarily the teacher will need to summarize scores or marks upon a comparatively small number of pupils; most often fifty or fewer. In such a case, long methods of calculation will not be found very uneconomical. On the other hand, any teacher must find a great many averages in the course of a school year. To know a few short-cuts, if these are easy to learn, will save much useless figuring.

The average (arithmetic mean) is ordinarily defined as: the sum of the scores (or other numbers) divided by the number of scores summed. The statistician usually represents a score or other number by the letter X, and the number of cases (scores) by the letter N. He also indicates the operation of summing by the Greek letter sigma (Σ). Using these letters, the formula for the arithmetic mean is $M = \frac{\Sigma X}{N}$. This formula is read: "The mean equals the sum of the scores divided by the number of scores." There is an X for each pupil in the class, and N represents the number of pupils (in the case of tests, scores for a class).

Three methods of calculating the arithmetic mean or average may be described briefly. These are:

1. The long method for ungrouped measures.
2. The short method for ungrouped measures.
3. The short method for grouped measures.

Table 76 gives the scores earned by a class of thirty pupils on an objective test. This table will serve to illustrate the "long" and "short" methods for ungrouped measures (scores).

TABLE 76

SCORES OF THIRTY PUPILS ON AN OBJECTIVE TEST AND THE CALCULATION OF THE ARITHMETIC MEAN BY LONG AND SHORT METHODS

(a)	(b)	(c)	
PUPIL	X (Score)	x'	
		+	−
1	81	1	
2	90	10	
3	65		15
4	77		3
5	89	9	
6	99	19	
7	70		10
8	65		15
9	83	3	
10	86	6	
11	75		5
12	91	11	
13	66		14
14	68		12
15	90	10	
16	88	8	
17	70		10
18	73		7
19	79		1
20	82	2	
21	91	11	
22	83	3	
23	68		12
24	51		29
25	40		40
26	86	6	
27	69		11
28	77		3
29	85	5	
30	97	17	
Σ (Sums)	(2334)[1]	121	−187
		−66	
Guessed Mean (M′) = 80			

COMPUTATIONS

Long Method:

$$M = \frac{\Sigma X}{N} = \frac{2334}{30} = 77.8$$

Short Method:

$$M = M' + \frac{\Sigma x'}{N}$$

$$= 80 + \frac{-66}{30}$$

$$= 80 - 2.2 = 77.8$$

[1]Note that this sum *need not* be found by the short method. It is inserted here merely to make this table show the procedure by both methods.

The long method for ungrouped measures. Column (*b*) of Table 76 shows the scores of the thirty pupils. The computational steps are:

1. Add the thirty measures. (Sum = 2334.) We can express this step as $\Sigma X = 2334$.

2. Divide the sum of the measures by N (30). The arithmetic mean (M) is therefore: $2334 \div 30 = 77.8$.

The short method for ungrouped measures. Table 76 also shows the computations for this method (especially column *c*). Note that we have introduced a new symbol, x'. The meaning of this will become clear in the outline of the steps employed in the short method.

1. Inspect the series of scores or numbers carefully. Select some value which *appears* (*without computation*) to be a reasonable guess or estimate of the average. Call this value M'. In Table 76, the value 80 was taken as M' (known variously as the "guessed average," the "assumed mean" or the "arbitrary origin").

2. Subtract each measure (X value) from the guessed mean (M'), algebraically, i.e., observing signs. (See column *c*.) It is often helpful, as was done in Table 76, to keep positive and negative values of these differences (x') in separate columns.

3. Add the x' values, obtaining their algebraic sum (-66 in our example).

4. Substitute in the following formula and solve:

$$M = M' + \frac{\Sigma x'}{N}$$
$$= 80 + \frac{-66}{30}$$
$$= 80 - 2.2 = 77.8 \text{ (arithmetic mean).}$$

It should be noted that the final value of the mean is the same by both long and short methods. This will always be found to be the case, as the short method *automatically* corrects the guessed mean (M'). It makes no difference whether

the guessed mean (M') is taken near the true mean (M) or whether we make a "bad" guess, the correction is made automatically. However, a poor guess does increase the size of the numbers to be handled and increases slightly the danger of a computational error. Since we are interested in saving labor, the attempt should be made to secure a fairly reasonable estimate at the outset, but, *spending much time on the selection of M' will defeat the sole purpose of the short method, viz., economy of time and labor.*

The short method does not always represent time saved. It usually is advantageous if twenty or more numbers must be handled. The principal advantage of the short method is that it gives us small numbers to deal with and hence reduces mental effort with consequent probability of fewer computational errors and re-checkings of calculations.

The short method for grouped measures. If a large number of measures (say, 50 to 100, or more) are to be averaged, statisticians often simplify their computations by what is called *grouping*. In other language they form a *frequency distribution* of the measures. Grouping is essentially the throwing together, as classes, measures which fall close together. We shall illustrate the method by the use of the same thirty scores of Table 76. The steps follow:

1. Find the range of scores (the difference between the highest and lowest measure). In this case the range is $99-40=59$.

2. Divide the range (here 59) by some number which will give a quotient of from 15 to 20. Three would be the best number in this case. We will call three the *range of the class-interval* and designate it by the symbol i. This means that we will group the thirty scores to the nearest three.

3. Form *class-intervals* by 3's, proceeding from the large values to the small thus:

98 to 100
95 to 97
92 to 94 Etc.

Taking the first mentioned class-interval (98 to 100) as an illustration, this procedure means that we will consider all scores of 98, 99, or 100 as falling in the same class. More specifically, we will consider scores of 98, 99, and 100 as *all falling at the mid-point* (99) of that class (98 to 100).

We usually choose classes so that the range of the class-interval (i) is 2, 3, 5, 10 (or some multiple of 5), when possible, *but the important thing is to be sure that the total number of classes does not fall much below, say, fifteen.*

4. Classify the measures falling into each class-interval as shown in Table 77. Note that we have used column (b) for tallying the frequencies (numbers) falling in each class-interval, and that column (c) collects the tallies as numbers for the sake of convenience in subsequent multiplications.

5. Guess or assume some average (M'). This was taken to fall in the class-interval, 74 to 76. In order to give some particular number as the value of M', the *mid-point* (75) of this class is taken. (This is equivalent to assuming that in this class, as in all others, all measures falling within that class lie exactly on the mid-point of the class.) Draw lines, as shown, to indicate the class-interval in which the mean was assumed to lie.

6. Calling the deviations of each class-interval from the class-interval in which the mean was assumed to lie (74 to 76) x', write in the deviations (x') both ways from the assumed mean, as shown in column (d). Give signs to these values to show whether the deviations are above or below the assumed mean. Note that the values of x' are in terms of class-intervals, not actual scores. Thus the class 71-to-73 is taken at -1, meaning that it is one class below the assumed mean. In reality, the measures falling in the class 71-to-73 are three scores below the class 74-to-76, in which the mean was assumed to lie, since the grouping was by 3's. The formula used for finding the arithmetic mean for grouped distributions automatically takes care of the fact that the stated x' values are but one-third of their true size.

TABLE 77

THE SCORES OF TABLE 76 RE-CLASSIFIED AS A FREQUENCY DISTRIBUTION USING 3 AS A CLASS-INTERVAL

(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
CLASS-INTERVAL	TALLY	f	x'	fx'
98 to 100	/	1	8	8
95 to 97	/	1	7	7
92 to 94			6	
89 to 91	////	5	5	25
86 to 88	///	3	4	12
83 to 85	///	3	3	9
80 to 82	//	2	2	4
77 to 79	///	3	1	3
74 to 76	/	1	0	(68)
71 to 73	/	1	−1	−1
68 to 70	////	5	−2	−10
65 to 67	///	3	−3	−9
62 to 64			−4	
59 to 61			−5	
56 to 58			−6	
53 to 55			−7	
50 to 52	/	1	−8	−8
47 to 49			−9	
44 to 46			−10	
41 to 43			−11	
38 to 40	/	1	−12	−12
Σ (Sums)		30		(−40)
				68
				−40
$M=M'+i\frac{\Sigma fx'}{N}=75+3\frac{28}{30}=77.8$				28

7. Multiply each x' by the corresponding f and record the products in the fx' column (column *e*).

8. Add the fx' column.

9. Substitute in the formula:

$$M=M'+i\frac{\Sigma fx'}{N}$$

$$=75+3\frac{28}{30}=75+2.8=77.8$$

It happened in this case that we obtained the same value for the mean as we did by the two preceding methods. This

exact correspondence was a matter of chance. Ordinarily there will be a slight disagreement between the means figured by grouped and ungrouped methods. The reader may test this by re-classifying the thirty scores with a grouping by 5's. As a rule, if N is fairly large and the number of class-intervals does not fall much below 15, the grouped distribution will yield a mean very close to the ungrouped value. A little thought will show that grouping will inevitably distort the actual values somewhat but that this distortion tends to cancel out in the long run. In any event, examination marks and test scores are at best estimates of the facts, and hence a slight distortion by grouping has little significance in comparison with the economy of the method. As was noted before, a distribution with but thirty cases is not adequate to show the real economies of grouping. Had there been 300 pupils, the saving in time and energy would have been very evident. Like the two preceding methods (for ungrouped measures), it does not matter particularly which class-interval is assumed to include the mean. A bad guess increases the computational labor, but the formula automatically corrects for the poor judgment.

PROBLEM 2

Determining the Reliability of a Test

This volume has made repeated reference to correlation or the mathematical relationship between two sets of test scores or other values. In particular, the terms *reliability coefficient* and *validity coefficient* have been used frequently without presenting any exact definitions. Because of the very great importance of correlation methods in test construction a brief introduction to the statistics of the measurement of relationship is given here. A concrete illustration is chosen as an introduction to the calculation of the coefficient of correlation.

As one assignment in a correspondence study course[1] the author outlines the following project:

1. Make up at least 20 or 25 broad essay-type or discussion questions (any subject). If possible, have each question call for something more than pure facts, i.e., call for some judgment, reasoning, originality, and organization of thought.

2. From the 20 or 25 questions select two sets of 10 questions *each* (20 in all). Make these two sets of questions as equal in difficulty as possible. *The aim here is to prepare two examinations* which are equally difficult, valid, and reliable so far as it is within your power to judge.

3. Call one set of questions "Examination A" and the other "Examination B." Give "A" one day and "B" the next day, both to the *same* pupils.

4. Erase the pupils' names before grading and replace the names by numbers. Then shuffle the papers and grade both sets upon the scale of 100 points. Etc.

Below are the results from this experiment on a class of twenty-two pupils.

Pupil.....	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Examination A..	90	88	72	53	62	64	86	57	65	87	87	91	83
Examination B..	79	69	75	47	62	67	72	40	47	62	90	78	63

Pupil.....	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	Mean
Examination A..	61	78	67	91	91	94	75	90	67	77.2
Examination B..	32	59	71	87	53	78	51	66	58	63.9

Table 78 shows the tabulation of the above scores together with the steps in the calculation of the coefficient of correlation. To distinguish the scores or marks on the two examinations, we have designated the A scores as X and the B scores as Y. Then x' and y' represent deviations of the X and Y from their respective guessed means (M'_x and M'_y).

The values given in the row of sums at the bottom of Table 78 furnish the necessary facts for solving for r, the coefficient of correlation. The solution follows Table 78.

[1]Education 312AB, University of California, "New-Type or Objective Examinations, Assignments 11-12." The actual results quoted were contributed by Mrs. Violet G. Prather, Wasco, California.

TABLE 78

THE CALCULATION OF THE COEFFICIENT OF CORRELATION

PUPIL	X (Exam. A)	Y (Exam. B)	x'	y'	x'^2	y'^2	$x'y'$
1	90	79	15	14	225	196	210
2	88	69	13	4	169	16	52
3	72	75	− 3	10	9	100	−30
4	53	47	−22	−18	484	324	396
5	62	62	−13	− 3	169	9	39
6	64	67	−11	2	121	4	−22
7	86	72	11	7	121	49	77
8	57	40	−18	−25	324	625	450
9	65	47	−10	−18	100	324	180
10	87	62	12	− 3	144	9	−36
11	87	90	12	25	144	625	300
12	91	78	16	13	256	169	208
13	83	63	8	− 2	64	4	−16
14	61	32	−14	−33	196	1089	462
15	78	59	3	− 6	9	36	−18
16	67	71	− 8	6	64	36	−48
17	91	87	16	22	256	484	352
18	91	53	16	−12	256	144	−192
19	94	78	19	13	361	169	247
20	75	51	0	−14	0	196	0
21	90	66	15	1	225	1	15
22	67	58	− 8	− 7	64	49	56
$N=22$							
M'_x	75						
M'_y		65					
Σ (Sums)			49	−24	3761	4658	2682

$$r=\frac{\frac{\Sigma x'y'}{N}-\left(\frac{\Sigma x'}{N}\cdot\frac{\Sigma y'}{N}\right)}{\sqrt{\frac{\Sigma x'^2}{N}-\left(\frac{\Sigma x'}{N}\right)^2}\sqrt{\frac{\Sigma y'^2}{N}-\left(\frac{\Sigma y'}{N}\right)^2}}$$

$$r=\frac{\frac{2682}{22}-\left(\frac{49}{22}\cdot\frac{-24}{22}\right)}{\sqrt{\frac{3761}{22}-\left(\frac{49}{22}\right)^2}\sqrt{\frac{4658}{22}-\left(\frac{-24}{22}\right)^2}}$$

$$r=.665+=.67$$

The correlation between Examinations A and B is therefore 0.67. Since these two duplicate examinations may be

regarded as two independent measures or samplings of the same thing, we call such a correlation coefficient a *reliability coefficient.* The particular formula employed in this solution is one form of what is called the Pearson product moment formula for correlation.

The reader will be interested to note that the calculations necessary for finding r also provides the needed data for the calculation of the arithmetic means as well. To find M_x and M_y, we need only substitute the values for the $\Sigma x'(49)$ and the $\Sigma y'(-24)$ in the formula previously given for the means for ungrouped scores by the "short method." Thus:

$$M_x = M'_x + \frac{\Sigma x'}{N} = 75 + \frac{49}{22} = 77.2$$

$$M_y = M'_y + \frac{\Sigma y'}{N} = 65 + \frac{-24}{22} = 63.9$$

Returning to the reliability coefficient, it should be pointed out that the exact procedure followed in our illustration applies strictly only to the case where two duplicate or equivalent examinations have been given to the *same* pupils. There are in all three common methods of finding reliability coefficients:

1. By correlation of the scores from duplicate or equivalent examinations administered to the same pupils (as in the foregoing discussion). This is ordinarily the most accurate and defensible method.

2. By splitting the results from a single examination into chance halves, correlating the half-scores, and "stepping up" the resulting coefficient of correlation by means of the Spearman-Brown prophecy formula (to be described later).

3. By repeating the same test or examination after an interval and correlating the results. This is often called the "re-testing coefficient of reliability." This method should never be employed when the first or second methods are possible. The procedure for this method is exactly like that shown in Table 78.

Since teachers usually give a single examination over a particular unit of school work, Method 2 (chance halves) is probably of the most general utility. For this reason we shall illustrate the method by a second actual example.[1]

In order to keep the illustration close to a real school situation, an actual examination from a seventh-grade class in physiology and hygiene has been selected. The original scoring of the teacher has been used in all the computations. The examination included these five questions:

1. What are the two most important aims of physiology?
2. State five rules of health that upper-grade children should know and practice.
3. Name the two classes of muscles and give examples of each kind.
4. Give two uses of the bones.
5. Name three digestive juices and tell what each does to the food.

Thirty pupils wrote on these questions, and each question was graded on a basis of twenty points. In order to find out just how reliable the results of this examination were, the scores earned on the odd-numbered items (questions 1, 3, and 5) were added separately from those of the even-numbered items (questions 2 and 4). This is known as "breaking the examination into chance halves." Since but five questions were given, the "halves" cannot be made to contain the same number of questions, it being necessary to place three questions in one of the chance "halves" and but two questions in the other. This will not make any very important difference in the correlation. Table 79 gives the data as tabulated from the thirty papers.

The next step in the solution of our problem is that of obtaining the correlation of the chance halves of the examination. The method of carrying out the actual computations is shown by the solution in Table 79. The odds will be designated by X and the evens by Y, and these scores have been tabulated in the columns so designated.

[1]Taken, with changes, from the *Improvement of the Written Examination*, pp. 132ff.

TABLE 79

TOTAL SCORES AND SCORES ON CHANCE HALVES OF AN EXAMINATION IN PHYSIOLOGY AND HYGIENE FOR A SEVENTH-GRADE CLASS OF THIRTY PUPILS

PUPIL	SCORES BY QUESTIONS					SCORE ON ODDS	SCORE ON EVENS	TOTAL SCORE
	I	II	III	IV	V			
1	10	16	10	5	5	25	21	46
2	10	20	20	10	15	45	30	75
3	10	20	20	10	10	40	30	70
4	10	20	10	15	10	30	35	65
5	10	20	10	10	10	30	30	60
6	20	20	20	10	17	57	30	87
7	10	20	20	20	17	47	40	87
8	20	20	20	8	15	55	28	83
9	18	20	20	15	20	58	35	93
10	10	20	20	10	10	40	30	70
11	20	16	10	15	17	47	31	78
12	20	20	10	0	0	30	20	50
13	20	20	20	3	13	53	23	76
14	18	18	20	10	10	48	28	76
15	10	16	20	10	15	45	26	71
16	12	10	10	10	17	39	20	59
17	10	20	20	14	15	45	34	79
18	20	20	20	18	20	60	38	98
19	10	20	10	7	8	28	27	55
20	15	20	10	20	20	45	40	85
21	15	20	20	13	5	40	33	73
22	18	20	20	18	18	56	38	94
23	10	20	15	10	10	35	30	65
24	20	18	15	15	8	43	33	76
25	10	16	20	10	15	45	26	71
26	20	20	20	20	17	57	40	97
27	20	15	10	8	14	44	23	67
28	18	10	20	15	7	45	25	70
29	20	20	10	20	10	40	40	80
30	18	20	12	18	10	40	38	78

The procedure from here on is identical with that of Table 78 where the scores from two different examinations were correlated. M'_x was taken as 45 and M'_y as 30. (See Table 80.)

The solution for r is given in detail at the top of the next page.

$$r=\frac{\frac{\Sigma x'y'}{N}-\left(\frac{\Sigma x'}{N}\cdot\frac{\Sigma y'}{N}\right)}{\sqrt{\frac{\Sigma x'^2}{N}-\left(\frac{\Sigma x'}{N}\right)^2}\sqrt{\frac{\Sigma y'^2}{N}-\left(\frac{\Sigma y'}{N}\right)^2}}$$

$$=\frac{\frac{595}{30}-\left(\frac{-38}{30}\cdot\frac{+22}{30}\right)}{\sqrt{\frac{2614}{30}-\left(\frac{-38}{30}\right)^2}\sqrt{\frac{1130}{30}-\left(\frac{+22}{30}\right)^2}}=0.37$$

The arithmetic means are:

$$M_x=M'_x+\frac{\Sigma x'}{N}=45+\left(\frac{-38}{30}\right)=45-1.3=43.7$$

$$M_y=M'_y+\frac{\Sigma y'}{N}=30+\left(\frac{22}{30}\right)=30+.7=30.7$$

Table 80 shows the data for the calculation of the reliability coefficient by the method of chance halves.

The reliability coefficient for the halves (odds vs. evens) was found to be 0.37. This is to be interpreted as the reliability of either *half* of the examination. What we need is the reliability of the *whole* examination, i.e., for the five questions.

To obtain r_{12} (meaning the correlation which would be expected between the whole examination actually given and a second but hypothetical examination of the same length which might be made up of five similar questions) from $r_{\frac{1}{2}\frac{1}{2}}$ (meaning the correlation of chance halves which was actually computed), we need only substitute the value 2 for n, and 0.37 for r in the following formula:

$$r_{nn}=\frac{nr}{1+(n-1)r}$$

This formula is variously known as "Brown's formula" and as the "Spearman prophecy formula," the latter name being, perhaps, preferable. The solution for our problem is:

$$r_{nn}=\frac{2(0.37)}{1+(2-1)(0.37)}=0.54$$

TABLE 80

CALCULATION OF THE RELIABILITY COEFFICIENT BY THE METHOD OF CHANCE HALVES

X (Odds)	Y (Evens)	x'	y'	x'^2	y'^2	$x'y'$
25	21	−20	− 9	400	81	180
45	30	0	0	0	0	0
40	30	− 5	0	25	0	0
30	35	−15	5	225	25	−75
30	30	−15	0	225	0	0
57	30	12	0	144	0	0
47	40	2	10	4	100	20
55	28	10	− 2	100	4	−20
58	35	13	5	169	25	65
40	30	− 5	0	25	0	0
47	31	2	1	4	1	2
30	20	−15	−10	225	100	150
53	23	8	− 7	64	49	−56
48	28	3	− 2	9	4	− 6
45	26	0	− 4	0	16	0
39	20	− 6	−10	36	100	60
45	34	0	4	0	16	0
60	38	15	8	225	64	120
28	27	−17	− 3	289	9	51
45	40	0	10	0	100	0
40	33	− 5	3	25	9	−15
56	38	11	8	121	64	88
35	30	−10	0	100	0	0
43	33	− 2	3	4	9	− 6
45	26	0	− 4	0	16	0
57	40	12	10	144	100	120
44	23	− 1	− 7	1	49	7
45	25	0	− 5	0	25	0
40	40	− 5	10	25	100	−50
40	38	− 5	8	25	64	−40
Σ (Sums).....		−38	+22	2614	1130	+595
Guessed Mean 45	30					

The value for n is taken as 2, since the whole examination is two times the length of the half examinations. It is obvious that such an examination is not very reliable; in fact, it should be considered quite inadequate for purposes of measurement. Pupils tested with this examination would

in many cases earn very different marks if the examination were repeated on another day with five questions as similar in type to these as possible, but not identical in the knowledge called for.

PROBLEM 3

Uses of the Spearman-Brown Prophecy Formula

We have already seen one use of the Spearman-Brown formula in connection with the physiology examination whose reliability for chance halves was found to be 0.37. "Stepping up" the 0.37 by the formula gave 0.54 as the *estimated* reliability of the entire examination of five questions. Many similar uses might be cited. Before doing this it should be noted that Table 81 is very useful in obtaining directly approximately accurate results for many values of r and n. If more exact results are needed, interpolation is necessary.[1]

We also found the examination of Table 78 to be .67 in reliability. We might wish to know the reliability of the sum or average of Examinations A and B. Examinations A and B are twice as long as either one alone; n is therefore 2. By direct substitution in the Spearman-Brown formula, setting r equal to .67 and n equal to 2, we obtain .80 as our estimate. Exactly the same value is obtained by interpolation between .60 and .70 in Table 81. The reader should note that the reliability of the sum and of the average of two examinations is always the same.

---

[1]Interpolation may be illustrated as follows: Suppose the reliability of the chance halves of an examination is 0.37 (as found above for the physiology examination). The value 0.37 does not occur in Table 79. We do find .30 and .40. Since n equals 2 when halves are "stepped up" to estimate the whole, we can interpolate as follows:

r	r_{nn} (when $n=2$)
.30	.57
.40	.46
	.11 (difference)

Since .37 is $\frac{7}{10}$ of the distance from .30 to .40, the value of r_{nn} must be (approximately) $\frac{7}{10}$ of the distance between .46 and .57, or $\frac{7}{10}$ of .11, or about .8. Adding .8 to .46 we have .54 as the estimated value of r_{nn} when $r_{\frac{1}{2}\frac{1}{2}}$ is .37. This agrees with the value previously found by actual substitution in the Spearman-Brown formula.

TABLE 81

TABLE FOR OBTAINING DIRECTLY VALUES OF r_{nn} FOR THE SPEARMAN-BROWN PROPHECY FORMULA FOR VARIOUS VALUES OF r AND n

r	n								
	2	3	4	5	6	7	8	9	10
.10	.18	.25	.31	.36	.40	.44	.47	.50	.53
.20	.33	.43	.50	.56	.60	.64	.67	.69	.71
.30	.46	.56	.63	.68	.72	.75	.77	.79	.81
.40	.57	.67	.73	.77	.80	.82	.84	.86	.87
.50	.67	.75	.80	.83	.86	.88	.89	.90	.91
.60	.75	.82	.86	.88	.90	.91	.92	.93	.94
.70	.82	.87	.90	.92	.93	.94	.95	.95	.96
.80	.89	.92	.94	.95	.96	.96	.97	.97	.98
.90	.947	.964	.973	.978	.981	.984	.986	.988	.989
.91	.953	.968	.976	.981	.984	.986	.988	.989	.990
.92	.958	.972	.979	.983	.986	.988	.989	.990	.991
.93	.964	.976	.982	.985	.988	.989	.991	.992	.993
.94	.969	.979	.984	.987	.989	.991	.992	.993	.994
.95	.974	.983	.987	.990	.991	.993	.993	.994	.995
.96	.980	.986	.990	.992	.993	.994	.995	.995	.996
.97	.985	.990	.992	.994	.995	.996	.996	.997	.997
.98	.990	.993	.995	.996	.997	.997	.997	.998	.998
.99	.995	.997	.997	.998	.998	.999	.999	.999	.999

Another use of the Spearman-Brown formula may be illustrated. A teacher gave a true-false test of 80 items. The reliability of odds vs. evens proved to be .60. From Table 81 it is evident that the reliability of the whole test was about .75. Since the reliability of the halves was .60, we may say that forty items (half the test) showed a reliability of .60. This teacher had hoped to obtain a reliability of .90 for her examination. The problem is to find how many items (of the same kind as she used) will be needed to yield a reliability of .90. Starting with $r = .60$, and reading along the row toward the right, we find .90 in the column headed 6. This teacher will therefore need about 6×40 items or 240 items to obtain the desired .90. This, of course, is an *estimate*. To be safe she should plan on at least 250 items.

Standard-test makers find the estimates yielded by the Spearman-Brown formula almost indispensible. The teacher who takes the trouble to calculate reliabilities on her more important tests will also find Table 81 of considerable value, at least for purposes of reference in her thinking.

Table 81 also shows very clearly that if an examination requiring a period (30 to 60 minutes) of class time yields a reliability of less than .50, it is almost hopeless to attempt to secure a reliable test by lengthening such an examination. It would have to be nine times as long to yield .90 as the reliability (if $r = .50$). An examination of the degree of reliability represented by .30 or .40 could never be lengthened within the time justified in actual teaching to become very satisfactory.

PROBLEM 4

The Measurement of Variability or Dispersion

It is well known that the average gives but a partial picture of the facts about a number of scores or marks. Two classes might have the same average while showing marked differences in the spread or scatter of the pupils about that average. We say that the two classes show a different variability or dispersion about the central tendency.

As a crude measure of variability, the range of the scores is sometimes used, thus:

Class	Average	Lowest Score	Highest Score	Range
A	81	34	112	78
B	81	69	99	30

Class A shows a range more than two and one-half times as great as Class B. Although the range is often of considerable value in expressing variability, it is open to the serious objection that it is too greatly influenced by occasional and chance factors which may give rise to one or more extreme deviates that are hardly typical. A single very bright or

very dull pupil may exert an enormous influence on the range. Consider the following facts:

Class	Average	Lowest Score	Second Lowest Score	Highest Score	Range
C	76	13	51	102	89
D	76	50	52	115	65

Class C has a much larger range (89) than Class D (65). If we examine these situations more closely, we see that the larger range of Class C is due to one extreme deviate (very low score), the pupil earning the score of 13. If we should take a number of similar classes of the same size, it is very unlikely that we should often find scores as low as 13. That pupil is atypical or unusual.

The usual measure of variability is the *standard deviation.* It is a more stable measure than the range and has the important advantage of not being unduly influenced by occasional and atypical, extreme variations. The standard deviation is also referred to as *sigma,* and is abbreviated to *S. D.,* or the Greek letter σ. The standard deviation requires considerable computation, although it is obtained easily as a by-product from the calculation of the coefficient of correlation. This will be clear from comparing the formulas below with Tables 78 and 79 and the subsequent calculations based upon these tables. We may give here four variations of the formula for the standard deviation, the choice among these four depending upon the needs in a particular instance, i.e., whether the measures are ungrouped or grouped and whether the deviations are taken from the true mean or from a guessed mean.

(1) $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ ..... { Ungrouped measures where the deviations (x) are taken from the true mean (M).

(2) $\sigma = i\sqrt{\frac{\Sigma x^2}{N}}$ ..... { Grouped measures where the deviations (x) are taken from the true mean (M).

(3) $\sigma = \sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2}$ ..... {Ungrouped measures where the deviations (x') are taken from a guessed mean (M').

(4) $\sigma = i\sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2}$ ..... {Grouped measures where the deviations (x') are taken from a guessed mean (M').

Formula (1) will serve best to define the standard deviation. We will now apply it to a very simple example, the X values being the marks of seven pupils.

TABLE 82

CALCULATION OF THE STANDARD DEVIATION.

X (Marks)	x	x^2	
93	8	64	$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$
88	3	9	
87	2	4	
85	0	0	$= \sqrt{\frac{152}{7}}$
84	−1	1	
80	−5	25	
78	−7	49	$= 4.66 = 4.7$
$M = 85$		$\Sigma = 152$	

The very simple example of Table 82 shows the essential nature of the standard deviation. We took the square root of the sum of the squares of the deviations, after dividing this sum by N. Some scientists, particularly astronomers and physicists, call the standard deviation the "root mean square deviation," meaning that the square root is taken of the mean (or average) of the deviations squared. A standard deviation based upon but seven cases is, of course, meaningless.

If we turn to the formula for r given in connection with Table 78, we see at once that the two radical terms in the denominator are identical with Formula (3) above; except, of course, for the use of y' instead of x' in the second term in order to distinguish the two sets of scores. All that is

necessary in order to calculate σ_x (the standard deviation of the X scores) and σ_y (the standard deviation of the Y scores) is to extract the indicated square roots, thus:

$$\sigma_x = \sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2} = \sqrt{\frac{3761}{22} - \left(\frac{49}{22}\right)^2} = 12.9$$

$$\sigma_y = \sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N}\right)^2} = \sqrt{\frac{4658}{22} - \left(\frac{-24}{22}\right)^2} = 14.5$$

Solving the two radical expressions in the denominator of the r formula for the data following Table 79, we obtain $\sigma_x = 9.2$ and $\sigma_y = 6.1$.

Formula (3) is therefore basic to the calculation of the coefficient of correlation. When actual correlations are computed, the standard deviations require only a small additional calculation (the taking of the square roots of the two radical terms of the denominator). This gives the standard deviation as a by-product of the correlation process in much the same way that the means (averages) are yielded incidentally in finding r. The teacher need not avoid the use of the standard deviation because of extra labor, since the standard deviation is ordinarily not needed unless a critical study of reliability is undertaken, in which case r_{xy}, M_x, M_y, σ_x, and σ_y are computed as one general calculation.

We should now note certain facts about the standard deviation. If we assume the distribution to be normal, the facts shown by Fig. 16 on page 426 hold.

Between the mean and either plus- or minus-one standard deviation, there are 34.13 per cent or slightly more than one third of the total cases in a normal distribution. For distributions which are roughly normal the same figures hold approximately. Beyond plus- or minus-one standard deviation there must be about one sixth of the total number of cases. Between plus-one sigma and minus-one sigma there are roughly two-thirds of the cases. Although these

facts hold, strictly speaking, only for normal distributions, most distributions of educational abilities for a random population will be found to be sufficiently normal to permit

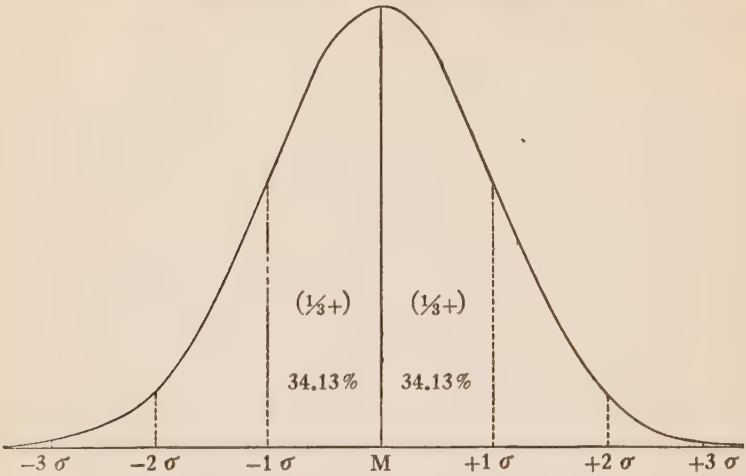


FIG. 16.—The standard deviation and the normal curve.

treatment as such. We can check these statements in a rough way for the two sets of marks given in Table 78.

	X (Exam. A)	*Y* (Exam. B)
Mean	77.2	63.9
Standard Deviation	12.9	14.5
$+1\sigma$	90.1 (77.2+12.9)	78.4 (63.9+14.5)
-1σ	64.3 (77.2−12.9)	49.4 (63.9−14.5)
Cases between $+1\sigma$ and -1σ	13 (By actual count)	15 (By actual count)

With a total population (N) of 22 cases, the most probable number falling within the limits $+1\ \sigma$ and $-1\ \sigma$ is about 15, i.e., $.6826 \times 22 = 15.0$. The actual cases included between $+1\ \sigma$ and $-1\ \sigma$ were 13 and 15, respectively. It must be

evident that the percentages stated above will hold but very roughly when N is small, as in this case.

Before closing the discussion of the standard deviation as a measure of variability, we should refer to a statistical measure which is commonly derived from the standard

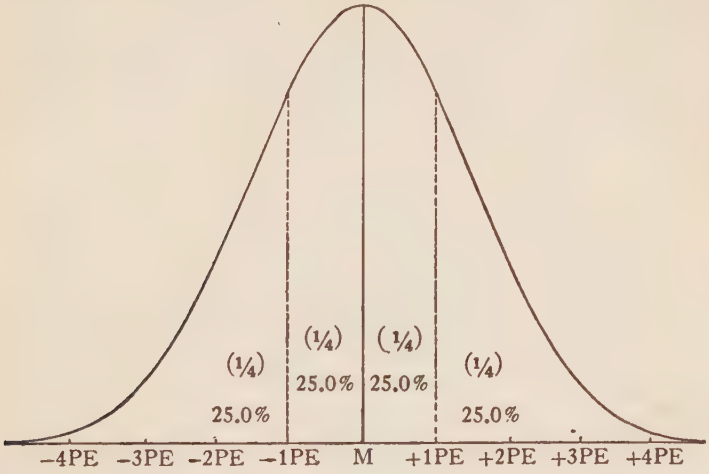


FIG. 17.—The probable error and the normal curve.

deviation. Ease of thinking about distributions would be fostered by having a measure of variability which would include between the mean and that measure exactly twenty-five per cent or one fourth of the cases in a normal distribution. Such a measure would divide the total distribution into quarters, fractions which are even more exact and simpler than those yielded by the standard deviation. Such a measure is the *probable error*. It is defined and calculated by means of the following formula: $PE = .6745\ \sigma$.

For Examinations A and B of Table 78, the probable errors are:

$PE_x = .6745 \times 12.9 = 8.7$ (approximately 9 points)
$PE_y = .6745 \times 14.5 = 9.8$ (approximately 10 points)

It will be noted that the range between $+1PE$ and $-1PE$ does not include even approximately one-half of the cases. This is due to two causes: (1) the small number of cases, and (2) absence of normality in the distributions.

We shall make use of the probable error (PE) in a later section. It is sufficient to remember that it is nothing more nor less than the standard deviation multiplied by .6745 in order to fraction the total distribution of a normal curve into exact quarters for the sake of ease of thinking. Fig. 17 illustrates the probable error.

PROBLEM 5

The Accuracy of an Individual Score

Having in mind the meaning of such statistical measures as the arithmetic mean, the standard deviation, the probable error, and the coefficient of correlation, we have the tools for attacking the very pertinent problem of the degree of confidence which may be placed in an individual score or mark. This problem is often stated as that of finding the "probable error of a score"—not to be confused with the probable error of the distribution as already defined.

We are now familiar with one method of approach to this question, viz., the evaluation of a test through the calculation of its reliability coefficient, as in Table 78. It will be recalled that the same twenty-two pupils were given two supposedly equally good examinations. The correlation of the two sets of marks was 0.67, a low figure. This coefficient of correlation (here a reliability coefficient) shows that neither examination is highly dependable. It does not tell us directly how much error is likely to be present in the mark of any particular pupils. Pupil No. 1 earned 90 on Examination A and 79 on Examination B. Which is correct, or rather, which is more nearly correct? This pupil cannot very well

be both a "90" and a "79" pupil in any absolute sense. One fact is clear from our past calculations, viz., that Examination B was about thirteen points harder on the average than was Examination A.

Turn to Table 78 and try subtracting thirteen points from each mark in the X column. Compare these differences with the values in the Y column. Do the disagreements then disappear? By no means. In addition to the irregularities caused by the inequality of difficulty of the two examinations, there are other differences to be reckoned with. These differences are due to various sorts of unreliability—sampling, subjectivity, etc.

Were the results from Table 78 to be used for grading pupils by some ranking process, the average difference of 13.3 points (between means) would have small significance. If these marks are taken to be per cents or values on the scale of 100, the systematic difference of 13.3 points *is* serious. The fluctuating differences, not corrected by subtracting 13.3 from the marks on Examination A, would disturb either form of marking (ranks or per cents); the difference of 13.3 on the average would disturb only when per cents or some equivalent system are employed. We can see these situations more clearly from Table 83 on page 431, which presents further facts about the two sets of marks given in Table 78.

In studying Table 83 it is necessary to keep in mind that the differences in the $X-Y$ column are of two kinds: (*a*) those arising from the fact that Examination A was 13.3 points easier than Examination B on the average, and (*b*) those arising from unreliability proper. The mean difference (13.3) agrees, of course, with the difference previously reported in the calculations based on Table 78, where M_x was found to be 77.2 and M_y to be 63.9. The standard deviation of the differences proved to be 11.3, which, when

multiplied by .6745, gave 7.6 as the probable error of the differences. The value 7.6 is *independent* of the influence of the average difference of 13.3 points.[1]

Table 83 yielded 8.0 as the standard deviation of a score and 5.4 as the probable error of a score. As a matter of fact, we obtained these values in a most round-about manner, one which is seldom used in practice. This was done intentionally in order that something of the significance of fluctuations of scores of the same pupils from one test to another might be brought out. The method of Table 83 should be regarded as a "long" method, chiefly of explanatory interest to us.

When the standard deviations (of the distributions) and the correlation of the two sets of marks or scores are known, as in the present instance, we can write formulas for the standard deviation or probable error of a score as follows:

$$\sigma_{(\text{Score})} = \frac{\sigma_x + \sigma_y}{2}\sqrt{1-r_{xy}}$$

$$PE_{(\text{Score})} = .6745\frac{\sigma_x + \sigma_y}{2}\sqrt{1-r_{xy}}$$

[1]The truth of this statement may not be apparent. The standard deviation of the differences was calculated by a formula not previously given, but, of course, algebraically equivalent to all four formulas given on pp. 423-4. If we take the differences of the $X-Y$ column and regard them as a new set of X or raw score values (not to be confused with the marks from Examination A, previously denoted by X), the final column of Table 83 may be regarded as X^2 values. The formula for the standard deviation becomes in this case

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

The second term under the radical is the same as M_x^2. The formula may therefore be rewritten

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - M_x^2}$$

This formula differs algebraically only from the four others previously listed. It should be evident that the present formula differs from the others merely in guessing the average at zero instead of some value nearer the truth. If the reader will study Table 83 carefully, it will be seen that the present formula is an algebraic variation of the others previously discussed.

To return to the original purpose of this footnote, it need only be pointed out that the 13.3 (or average difference in difficulty of the two examinations) is eliminated by the second term under the radical, viz., $\left(\frac{-293}{22}\right)^2$.

Due to the fact that the two standard deviations are likely to differ somewhat, the formula calls for averaging the two.

Substituting in the foregoing formulas, we obtain:

$$\sigma_{(\text{Score})} = \frac{12.9+14.5}{2}\sqrt{1-.67} = 7.9$$

$$PE_{(\text{Score})} = .6745\,\frac{12.9+14.5}{2}\sqrt{1-.67} = 5.3$$

TABLE 83

CALCULATION OF THE PROBABLE ERROR OF A TEST SCORE
(Data From Table 78)

PUPIL	X (Exam. A)	Y (Exam. B)	DIFFERENCES $(X-Y)$	SQUARES OF DIFFERENCES
1	90	79	−11	121
2	88	69	−19	361
3	72	75	3	9
4	53	47	− 6	36
5	62	62	0	0
6	64	67	3	9
7	86	72	−14	196
8	57	40	−17	289
9	65	47	−18	324
10	87	62	−25	625
11	87	90	3	9
12	91	78	−13	169
13	83	63	−20	400
14	61	32	−29	841
15	78	59	−19	361
16	67	71	4	16
17	91	87	− 4	16
18	91	53	−38	1444
19	94	78	−16	256
20	75	51	−24	576
21	90	66	−24	576
22	67	58	− 9	81
Σ (Sums)			−293	6715

$$M_{(\text{Differences})} = -293 \div 22 = 13.3$$

$$\sigma_{(\text{Differences})} = \sqrt{\frac{6715}{22} - \left(\frac{-293}{22}\right)^2} = 11.3$$

$$PE_{(\text{Differences})} = .6745 \times 11.3 = 7.6$$

$$\sigma_{(\text{Score})} = .707\,\sigma_{(\text{Differences})} = .707 \times 11.3 = 7.99 = 8.0$$

$$PE_{(\text{Score})} = .6745\ \sigma_{(\text{Score})} = .6745 \times 8.0 = 5.4$$

The values given in the formulas at the top of page 431 differ but one point in the first decimal place from those calculated in Table 83 by the "long" method. The two methods may disagree slightly in a particular instance, for reasons which will not be discussed here.

We will consider the probable error of a score to be about 5.3 points.[1] The next task before us is to understand the meaning of such probable errors. To do this, consider the errors in a series of test scores for a large number of pupils to be distributed in a normal fashion. If the actual errors were plotted as a normal distribution, the probable error of that distribution would be 5.3 points. This again means that in fifty per cent of the cases the errors would fall between -1 *PE* and $+1$ *PE*, i.e., within a range of $5.3+5.3$ or 10.6 points. In other words, it is an even break or chance that any score is within 10.6 points of the correct score, were the latter obtainable. If a pupil took a very large number of tests similar to Examinations A and B, the average of all his marks would give his true score (if we neglect practice effects). The chances are therefore 50:50 that an obtained score is within 5.3 either way of the true score.

A probable error of 5.3 points is rather disconcertingly large when we recall that half of the pupils will receive scores with errors larger than this value. It can be shown that the following statements are true:

The chances are one to one that an obtained score is in error by not more than 5.3 points either up or down.

The chances are four to one that an obtained score is in error by not more than 10.6 (2 *PE*) points either up or down.

[1]The formula used here is not, theoretically, the best one to employ in case a highly accurate determination of the probable error of a score is needed. It ignores for one thing the fact that scores in different parts of the total range do not have equal possibilities for error. Very high scores are somewhat more likely to be in error upwards, i.e., to be too high. The converse is also true. This is the regression effect already mentioned. A score near the mean will have a somewhat smaller probable error than will an extreme deviate. The probable error formula adopted here may be thought of as a sort of "average probable error of a score." For further information on this point see T. L. Kelley, *Statistical Method*, (Macmillan, 1923), pp. 214ff; A. S. Otis, *Statistical Method in Educational Measurement*, (World Book Co., 1925), pp. 247ff; and G. M. Ruch, "Minimum Essentials in Reporting Data on Standard Tests," *Journal of Educational Research*, Vol. XII (1925), pp. 349-358.

The chances are twenty to one that an obtained score is in error by not more than 15.9 (3 PE) points either up or down.

The statements just given show very clearly that with a test of a reliability in the neighborhood of .65, there is really no assurance that a pupil earning 75 is really not a "65" or an "85" pupil. If his mark was 75, it is an even chance that he deserves a mark between 70 and 80. The chances are about four to one that he deserves a mark between 65 and 85. There is about one chance in twenty that he really should receive a mark as low as 65 or as high as 90. Since much evidence has been gathered by Monroe, the author, and many others to the effect that the usual examination falls very close to the value 0.65 in reliability, our present findings for the data of Tables 78 and 83 can hardly exaggerate the facts. To attempt to classify pupils in as small a number of groups as five (e.g., A, B, C, D, and E) would be "shaky business" with such a test, as a little figuring will show. The range of marks on Examination A of Table 78 was from 53 to 94, or forty-one points. For Examination B the range was from 32 to 90, or fifty-eight points. The average range was therefore not far from fifty points. If the $PE_{(\text{Score})}$ is five points and the range is fifty points, the following statements statements hold approximately:

The chances are one to one that an obtained score is in error by no more than one-fifth of the range.

The chances are four to one that an obtained score is in error by no more than two-fifths of the range.

The chances are twenty to one that an obtained score is in error by no more than three-fifths of the range.

PROBLEM 6

What is a Satisfactory Degree of Reliability?

The question of what constitutes a satisfactory degree of reliability cannot be answered in other than relative terms. Assuming that the teacher's usual interest in reliability

coefficients centers about the reliability of a test given in a single class or grade, the following statements can probably be defended:[1]

Reliability Coefficient	Interpretation or Significance
.95 to .99	Very high; rarely obtained except with long, carefully standardized tests. Long objective tests occasionally reach this level.
.90 to .94	High; about the limit of teacher-made tests of (say) 100 to 200 items.
.80 to .89	Fairly high; usually obtainable with well-constructed objective tests of 75 to 150 items. Relatively few essay-type examinations reach this level.
.70 to .79	Rather low; not of much value for purposes of evaluating individual pupils.
Below .70	Low; almost valueless except for averages of classes. The average essay-type examination does not exceed .70 and is possibly lower.

The foregoing statements must be taken with due caution, since there are many factors to be considered. Many will think the interpretations presented to be very conservative. Evidence to be brought forward later will show that the reverse is probably true. The older textbooks on statistical methods have often tended to create the impression that correlations above .50 or .60 were at least "moderately high," and that those above .75 to .80 were "very high." This view is possibly partly to be explained upon the basis of a tendency to regard coefficients of correlation as *per cents*. Nothing could be farther from the truth.

The reader may wonder whether there is any basis at all for answering the question, "When is a correlation high?"

[1]The confining of the discussion to a single grade or class was done intentionally. It is a well-known fact that the same test will show very different reliability coefficients in a "narrow" group like a single class in comparison with a "wide" group like one composed of a half dozen school grades pooled. We sometimes say that the correlation is dependent upon the range of talent or heterogeneity of the group. Kelley has given us a formula for inferring correlation for a wide range from a known value on a small range, or vice versa. Brief discussion of this point will be made later in the chapter.

Correlations are employed for two general, although not distinctly different, purposes: (*a*) as measures of relationship, and (*b*) for prediction. It is with the latter use that we need to deal in approaching a basis for interpreting correlation coefficients. We may as well return to the data of Table 78 for our illustration. We have previously obtained the following statistical constants:

M_x	M_y	SD_x	SD_y	r_{xy}
77.2	63.9	12.9	14.5	.67

In Table 78 we have given two sets of scores or marks as actually obtained on Examinations A and B. We can think of these as experimental or actual values. (They must not be regarded as *true* values.) To the degree that these two sets of scores are correlated, it is possible to predict one set of scores from the other.

If the correlation between the two sets of scores had been perfect (1.00 instead of .67), there would be no error in predicting the scores. If the correlation had been zero, the errors of prediction would be a maximum.

The actual formulas used for prediction are:

$$\overline{X} = r_{xy}\frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x$$

$$\overline{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}(X - M_x) + M_y$$

The bars above the $\overline{X}$ and $\overline{Y}$ are used to indicate that these are estimated or predicted values, not the obtained or actual experimental values.

If we substitute the statistical constants found for Table 78, we obtain the following:

$$\overline{X} = .67\,\frac{12.9}{14.5}\,(Y - 63.9) + 77.2, \text{ or } \overline{X} = .60Y + 39.1$$

$$\overline{Y} = .67\,\frac{14.5}{12.9}\,(X - 77.2) + 63.9, \text{ or } \overline{Y} = .75X + 5.8$$

These are called the *regression equations.* It should be noted that there are always two such regression equations, and that they are not interchangeable.

There are a great many educational situations where prediction by means of regression lines is possible and advantageous, but it must always be remembered that predicting scores or other values is inferior to actually obtaining such experimentally. Thus, it might have happened that, in addition to the twenty-two pupils of Table 78, certain additional pupils were absent and missed one or the other of the examinations. Instead of giving a later examination, the missing marks might have been predicted by the foregoing equations. It does not follow that the predicted mark would have been exactly the same one which the pupil would actually have earned. The higher the correlation, the safer the prediction method. Teachers often administer prognosis or prediction tests for purposes of sectioning classes or counseling pupils. Such tests have previously been proved to have such predictive powers. They are successful in practice only to the degree that their scores correlate with future success. Regression equations are often used to establish the best basis for such predictions.

Purely as a definition and illustration of prediction, we shall apply the regression equation to the values in the X column of Table 78; i.e., we shall predict the $\overline{Y}$ values which correspond to obtained or actual X values. There is no practical reason for doing this, except for illustration, since the actual Y values are known. However, comparing $\overline{Y}$ and Y values will give us a very good basis for arriving at some notions about the predictive value of a correlation of .67. Table 84 shows the results. The X column repeats the X column of Table 78. The Y column repeats the Y column of Table 78. The $\overline{Y}$ column shows the $\overline{Y}$ (predicted) scores for each X score in turn; the predicted scores being obtained by substituting each successive X value in the

TABLE 84

LONG METHOD OF CALCULATING THE STANDARD ERROR OF ESTIMATE

X (Actual scores on Exam. A)	Y (Actual scores on Exam. B)	$\overline{Y}$ (Estimated Scores on Exam. B by the equation, $\overline{Y}=.75X+5.8$)	$(Y-\overline{Y})$ (Errors of estimate)	$(Y-\overline{Y})^2$ (Squared errors of estimate)
90	79	73.3	− 5.7	32.49
88	69	71.8	2.8	7.84
72	75	59.8	−15.2	231.04
53	47	45.5	− 1.5	2.25
62	62	52.3	− 9.7	94.09
64	67	53.8	−13.2	174.24
86	72	70.3	− 1.7	2.89
57	40	48.5	8.5	72.25
65	47	54.5	7.5	56.25
87	62	71.0	9.0	81.00
87	90	71.0	−19.0	361.00
91	78	74.0	− 4.0	16.00
83	63	68.0	5.0	25.00
61	32	51.5	19.5	380.25
78	59	64.3	5.3	28.09
67	71	56.0	−15.0	225.00
91	87	74.0	−13.0	169.00
91	53	74.0	21.0	441.00
94	78	76.3	− 1.7	2.89
75	51	62.0	11.0	121.00
90	66	73.3	7.3	53.29
67	58	56.0	− 2.0	4.00

$\Sigma = 2580.86$

$$\sigma_{y.x}=\sqrt{\frac{2580.86}{22}}=10.8$$ (Standard Error of Estimate)

equation, $\overline{Y}=.75X+5.8$. The $Y-\overline{Y}$ column shows the differences between the actual Y and the predicted $\overline{Y}$ values. The $(Y-\overline{Y})^2$ column shows the squares of these differences.

The $Y-\overline{Y}$ values are termed the *errors of estimate.* By squaring these errors of estimates, taking their sum, dividing this sum by N, and finally extracting the square root, we obtain the *standard error of estimate.* A more descriptive, but less common, name would be the "standard deviation of

the errors of estimate" (or prediction). The symbol, $\sigma_{y.x}$, is used to indicate the standard error of estimate (of the Y's from the X's). Note the calculation of the value of $\sigma_{y.x}$ at the bottom of Table 84.

In order to illustrate the method of obtaining the $\overline{Y}$ values of Table 84, the actual solutions are given below for the first three pupils.

Pupil	X Score	Substitution in Equation	$\overline{Y}$ or Estimated Score
1	90	$\overline{Y}=(.75\times 90)+5.8$	73.3
2	88	$\overline{Y}=(.75\times 88)+5.8$	71.8
3	72	$\overline{Y}=(.75\times 72)+5.8$	59.8

Returning to the solution for the standard error of estimate at the bottom of Table 84, we can write the formula as follows:

$$\sigma_{y.x}=\sqrt{\frac{\Sigma(Y-\overline{Y})^2}{N}}.$$

It was suggested previously that this "long method" of finding the standard error of estimate was chosen merely for purposes of definition of such concepts as estimated scores, errors of estimates, etc. *In actual practice it is quite unnecessary to go through the laborious calculations of Table 84. The standard error of estimate is given directly* (for the estimation of $\overline{Y}$ values from X values) *by the following:*

$$\sigma_{y.x}=\sigma_y\sqrt{1-r_{xy}^2}.$$

Similarly, the standard error of estimating $\overline{X}$ values from Y values is:

$$\sigma_{x.y}=\sigma_x\sqrt{1-r_{xy}^2}.$$

If we substitute in the formula, $\sigma_{y.x}=\sigma_y\sqrt{1-r_{xy}^2}$, the actual values that we found previously for Table 78, we

obtain $\sigma_{y.x} = 14.5\sqrt{1-(.67)^2} = 10.7$ as the value of the standard error of estimate. This agrees closely with the result by the "long" method.

We are now ready to apply what we have learned about the standard error of estimate to our original problem of finding a basis of interpreting the meaning of correlation. In the formula, $\sigma_{y.x} = \sigma_y\sqrt{1-r_{xy}^2}$, r must take values between zero and 1.00 (unity). The sign of r can be either positive or negative, but this does not affect the degree of prediction. If r is 1.00, the radical expression reduces to zero, and hence the value of the standard error of estimate ($\sigma_{y.x}$) is zero. The prediction is perfect, since there is no error of prediction when r equals 1.00. Now note the other limiting value of r, viz., zero. When r is zero, the radical expression reduces to 1.00, and $\sigma_{y.x}$ equals σ_y. This shows that the standard error of estimate is exactly the same size as the standard deviation of the Y values when r is zero. There would be no guarantee at all that a pupil earning a very high score on Examination A (say, 95) would earn a high value on Examination B. In the absence of correlation, any value on Examination B would be as probable as any other. Since with zero correlation the errors of estimate are limited only by the variability of the Y scores, and no value of Y is more probable than any other for any given value of X, it may be said that zero correlation, if used for purposes of prediction, results in accuracy no better than chance.[1]

In our particular problem we found that with a correlation of .67 the standard error of estimate was 10.7, in comparison with the standard deviation of 14.5 for the Y values. The ratio of $\frac{10.7}{14.5}$ is .74; or in other words, the variability of our errors of estimate showed a spread roughly three-fourths as great as the scores being estimated. In a certain sense, therefore, the correlation of .67 is but one-fourth of

[1] *See* T. L. Kelley, *Statistical Method* (New York: The Macmillan Company, 1923), pp. 172-4.

the way (in accuracy of prediction) from zero correlation to perfect correlation.

Since the value of the standard error of estimate is always some fraction of that of the standard deviation, the fraction being given by the radical expression $\sqrt{1-r^2}$, it is possible to build up a simple table of great value in interpreting the significance of r for prediction. Kelley has termed this radical expression the *coefficient of alienation* (referring to the alienation or departure of any given value of r from 1.00 or perfect prediction).

Table 85 shows the coefficients of alienation and the per cent of reduction in the standard error of estimate for many values of r.

TABLE 85

COEFFICIENTS OF ALIENATION AND THE REDUCTION IN THE STANDARD ERRORS OF ESTIMATES FOR VARIOUS VALUES OF r

(*a*)	(*b*)	(*c*)
r	$\sqrt{1-r^2}$ COEFFICIENT OF ALIENATION	PER CENT OF REDUCTION IN STANDARD ERROR OF ESTIMATE
.00	1.000	0.0
.10	.995	0.5
.20	.980	2.0
.30	.954	4.6
.40	.917	8.3
.50	.866	13.4
.60	.800	20.0
.70	.714	28.6
.80	.600	40.0
.866+	.500	50.0
.90	.436	56.4
.95	.312	68.8
.96	.280	72.0
.97	.243	75.7
.98	.199	80.1
.99	.141	85.9
1.00	.000	100.0

Table 85 gives us some rather definite ideas about the accuracy of prediction possible with varying values of the

correlation coefficient. For example, it takes a value of r equal to about .866 to reduce the standard error of estimate to .50 (half of the standard deviation). When r is .95, column (*b*) shows that the alienation from perfect prediction is still almost one-third. In general it should be noted that the rise in accuracy of prediction is very slow for small values of r (.00 to .50 or .60); the rise shows positive acceleration, becoming very rapid only after .95 is passed. Even a correlation of .99 shows an alienation of about one-seventh from perfect prediction.

If the reader will now refer to the introductory statements about what constitutes high correlation, it will be granted that those statements were conservative indeed in the light of the mathematics of prediction.

PROBLEM 7

The Effect of Heterogeneity or Range of Talent on Reliability Coefficients

There is one additional concept which the student needs in considering the significance of correlations, particularly reliability coefficients. We can introduce the discussion by saying: *When correlation exists between two sets of variables (scores, marks, etc.), the amount of correlation is a function (result) of the variability of the group sampled.* This will be made clearer by the following example:

A teacher prepared an objective test in United States history. She gave this to her classes in the sixth, seventh, and eighth grades. She computed the reliability coefficients for each grade separately and for the three grades pooled. For the eighth grade alone, r was found to be .77; for the three grades pooled it was .91. This appears to be a contradiction. As a matter of fact, it is the expectancy. The more grades pooled, i.e., the greater the range of talent or abilities, the larger the expected correlation, other factors being the same.

In order to test out whether these two values of r, .77 and .91, are contradictory, we need certain additional facts, principally some measures of the relative variabilities of the two groups. The standard deviations are again best for our purposes.

Grade Range Tested	Reliability Coefficient	Standard Deviation
8th grade only.....	.77 (r)	10.2 (σ)
6th, 7th, and 8th grades....	.91 (R)	15.3 (Σ)

Note that certain letters have been inserted parenthetically after the values for the reliability coefficients and the standard deviations. These letters refer to the following formula:

$$\frac{\sigma}{\Sigma}=\frac{\sqrt{1-R}}{\sqrt{1-r}}.$$

Small letters refer to the statistical constants for the narrow range of talent (8th grade), and large letters indicate corresponding values for the broader range of talent (three grades pooled). It should be noted that Σ here means the standard deviation of the wider range of talent, 15.3, and not summation, as we have previously used it.

If we substitute for r, σ, and Σ the actual values obtained, and then solve the expression, we find a predicted value for R, thus:

$$\frac{10.2}{15.3}=\frac{\sqrt{1-R}}{\sqrt{1-.77}}. \quad \text{From which, } R=.90.$$

This means that the increase in the variability of the group tested from a standard deviation of 10.2 to a standard deviation of 15.3 should increase the reliability from .77 to about .90, provided that the test functions equally reliably throughout such a range of three grades. Since the .90 agrees very well with the actual value, .91, it may be concluded that the two sets of results are quite in harmony, the apparent discrepancy being explainable upon a basis of the

difference in the range of abilities normally found in a single grade and those found over a range of two or more grades.

A reliability coefficient is thus seen to be a function of the standard deviation. *For this reason the statement that a certain test has a reliability of .65 or .89 is quite meaningless unless additional facts are given.* In addition, some measure of heterogeneity or range of talent is needed. In rough work it is often meaningful to express range of talent in terms of the grades or ages of pupils tested. But, as our formula implies, the best measure of heterogeneity is the standard deviation. It follows that a reliability coefficient should be accompanied by a statement of the standard deviation. This entails no extra work, as we have seen, since the standard deviation is given as a by-product of the computation of the reliability coefficient.

The reader should be reminded at this point of the discussion under Problem 5, The Accuracy of an Individual Score. A formula was given for the probable error of a score, as follows:

$$PE_{\text{(Score)}} = .6745\frac{\sigma_x + \sigma_y}{2}\sqrt{1-r_{xy}}.$$

It should be noted that this formula includes both the standard deviation and the reliability coefficient. If the range of talent is increased (e.g., several grades instead of one), the standard deviations become larger, but the reliability coefficient also increases in size. If all other factors are constant, these two measures vary together. The $PE_{\text{(Score)}}$ is therefore roughly constant. For this reason it is usually more meaningful than the reliability coefficient itself.

If we compute the $PE_{\text{(Score)}}$ for both narrow and wide ranges of talent for the history test we have been using as an example, we obtain:

(Narrow Range) $PE_{\text{(Score)}} = (.6745)(10.2)\sqrt{1-.77} = 3.0$

(Wider Range) $PE_{\text{(Score)}} = (.6745)(15.3)\sqrt{1-.91} = 3.1$

TABLE 86

VALUES OF R CORRESPONDING TO VARIOUS VALUES OF r AND $\frac{\sigma}{\Sigma}$ IN THE FORMULA $\frac{\sigma}{\Sigma} = \frac{\sqrt{1-R}}{\sqrt{1-r}}$

$\frac{\sigma}{\Sigma}$	.00	.10	.20	$\frac{r}{.30}$	.40	.50	.60	.70	.80	.90	1.00
.10	.990	.991	.992	.993	.994	.995	.996	.997	.998	.999	1.000
.20	.960	.964	.968	.972	.976	.980	.984	.988	.992	.996	1.000
.30	.910	.919	.928	.937	.946	.955	.964	.973	.982	.991	1.000
.40	.840	.856	.872	.888	.904	.920	.936	.952	.968	.984	1.000
.50	.750	.775	.800	.825	.850	.875	.900	.925	.950	.975	1.000
.60	.640	.676	.712	.748	.784	.820	.856	.892	.928	.964	1.000
.70	.510	.559	.608	.657	.706	.755	.804	.853	.902	.951	1.000
.80	.360	.424	.488	.552	.616	.680	.744	.808	.872	.936	1.000
.90	.190	.271	.352	.433	.514	.595	.676	.757	.838	.919	1.000
1.00	.000	.100	.200	.300	.400	.500	.600	.700	.800	.900	1.000

These probable errors of score are practically constant, although the correlation was raised from .77 to .91 by pooling three grades. The slight disagreement is due to the fact that both the .77 and the .91 contain errors of sampling, since both are based upon rather small numbers of cases.[1]

Table 86 will be found convenient for reading directly values of reliability coefficients (R) for a wider range of talent or heterogeneity when two facts are given: the reliability on a narrow range (r) and the ratio of the standard deviations for the two ranges $\left(\frac{\sigma}{\Sigma}\right)$. Simple interpolation will be satisfactory for intermediate values.

[1]In the formula for the PE (Score) as used here, the average of the two standard deviations is not taken since σ_x is assumed to be equal to σ_y. (See Problem 5 of this chapter.)

GENERAL BIBLIOGRAPHY

The following bibliography was assembled by Mr. Sanford Siegrist and the author, Mr. George Meyer being responsible for many of the annotations and for general checking of the references. Space considerations led to the abandonment of the original intention of annotating all titles. The more important references are given short characterizations. Titles referred to, discussed, or abstracted in the body of the text are indicated by the letter T in parentheses. A few references, not available in local libraries, have been taken from other bibliographies. This list of titles lays no claim to completeness, although it is thought that most of the more significant literature is covered. Due to the use of a classification with subheadings there are a number of duplications, particularly in the case of books or longer papers treating a number of phases of the subject.

I. Books, Monographs, and Bulletins

1. Ballard, P. B., *The New Examiner*. London, Hodder and Stoughton, 1924. 266 pp. (T).
   An unusually well written and early discussion of old- and new-type examinations, with special attention to the examination as a measuring instrument.
2. Brinkley, S. G., "Values of New-Type Examinations in the High School with Special Reference to History," Teachers College, Columbia University, *Contributions to Education*, No. 161, 1924. 121 pp. (T).
   An experimental study of essay and objective tests over the same ground in history, tending to prove that neither type is measurably superior under comparable conditions.
3. Bursch, J. F. and Meltzer, H. M., "The New Examination," School of Vocational Education, Oregon Agricultural College, Corvallis, Oregon, Bulletin No. 422, Sept., 1926. 40 pp.
   Discusses the characteristics of a good examination and gives a number of sample tests.
4. Ellis, R. E., *Standardizing Teachers' Examinations and the Distribution of Class Marks*. Bloomington, Illinois, The Public School Publishing Co., 1927. 170 pp. (T).
   A small volume largely devoted to the normal curve idea in marking. The approach is logical rather than experimental.
5. Greene, C. E., "New Type Tests," *Research Monograph No. 3*, Public Schools, Denver, Colo., 1926. 35 pp.
   A short but excellent collection of classroom tests illustrating rules for constructing such measures.
6. Hopkins, L. T., "The Construction and Use of Objective Examinations," Boulder Colo., University of Colorado, 1926. 119 pp.
   A collection of examinations made by one hundred and eighteen teachers in a course on new-type examination construction.

7. MILLER, G. F., "Objective Tests in High-School Subjects," Norman, Oklahoma. Published by the author, 1926. 168 pp.
A very good collection of objective tests.

8. MONROE, W. S. AND SOUDERS, L. B., "Present Status of Written Examinations and Suggestions for Their Improvement," Bureau of Educational Research, University of Illinois, Bulletin No. 17. Urbana, University of Illinois, 1923. 77 pp. (T).
Presents experimental studies of the reliabilities of written examinations and samples of objective examinations.

9. MONROE, W. S., "Written Examinations and their Improvement," Bureau of Educational Research, University of Illinois, Bulletin No. 9. Urbana, University of Illinois, 1922. 71 pp. (Out of print).
One of the earliest treatments of the relative merits of traditional and objective examinations.

10. MONROE, W. S. AND CARTER, R. E., "The Use of Different Kinds of Thought Questions in Secondary Schools and Their Relative Difficulty for Students," Bureau of Educational Research, University of Illinois, Bulletin No. 34. Urbana, University of Illinois, 1923. 26 pp.
Twenty kinds of thought questions are described.

11. ODELL, C. W., "Objective Measurement of Information," Bureau of Educational Research, University of Illinois, Circular No. 44, May 11, 1926. 27 pp.
Samples of fifteen types of objective test exercises.

12. ODELL, C. W., *Traditional Examinations and New-Type Tests*, Century Company, 1928. 469 pp. (T).
One of the best and most comprehensive treatments of school examination methods. This volume maintains a highly commendable balance of judgment. There is a carefully selected bibliography of one hundred titles.

13. ORLEANS, J. S. AND SEALY, G. A., *Objective Tests*, World Book Company, 1928. 373 pp.
The latest treatment of the new-type examination. Essentially an account of the building and application of a series of objective tests for Lewis County, New York, but of general interest.

14. PATERSON, D. G., *The Preparation and Use of New-Type Examinations*, World Book Company, 1925. 87 pp.
A very brief treatment of particular interest to high-school and college teachers.

15. RUCH, G. M. AND STODDARD, G. D., *Tests and Measurements in High School Instruction*, World Book Co., 1927; especially pp. 251-297. (T).
A treatment of the standard tests for secondary schools but containing a section on objective testing.

16. RUCH, G. M., *The Improvement of the Written Examination*, Scott, Foresman and Co., 1924. 193 pp. (T).
The first book exclusively on objective testing published in America.

17. RUCH, G. M. et al, *Objective Examination Methods in the Social Studies*, Scott, Foresman and Co., 1926. 116 pp. (T).
A series of experimental monographs on moot questions in objective examination methods. Although dealing specifically with history testing, the findings should hold for all types of factual measurement.

18. RUSSELL, C., *Classroom Tests*, Ginn and Co., 1926. 346 pp.
A very elementary but full discussion of the building of a series of objective tests by a group of Toledo, Ohio, teachers. Very little critical or experimental material.

19. STRICKLAND, V. L., "Objective Tests," Kansas State Agricultural College, Manhattan, Kansas, Bulletin XVIII, No. 2. 23 pp.

20. SYMONDS, P. M., *Measurement in Secondary Education*, The Macmillan Co., 1927; especially pp. 10-52 and 498-542. (T).
An excellent treatment of standard tests for high schools. The sections cited treat of objective tests, teachers marks, limitations of the traditional examination, etc.

21. TOOPS, H. A., "Trade Tests in Education," Teachers College, Columbia University, *Contributions to Education*, No. 115, 1921. 118 pp. (T).
One of the earliest experimental studies on recall, multiple-response, and true-false tests. A model for much of the later investigation.

22. WEIDEMANN, C. C., "How to Construct the True-False Examination," Teachers College, Columbia University, *Contributions to Education*, No. 225, 1926. 118 pp.
The most comprehensive study of the true-false test yet available. Contains considerable experimental work.

23. WOOD, B. D., *Measurement in Higher Education*, World Book Co., 1923. 337 pp.
The pioneer work on measurement on the college level, both informal objective tests and standard intelligence examinations. Contains an account of the earlier work on the new-type test at Columbia College.

24. WOOD, B. D., *New York Experiments with New-Type Modern Language Tests*, The Macmillan Co., 1927. 339 pp.
An excellent example of the critical study of the techniques and results of objective measurement in modern languages.

II. UNRELIABILITY OF TEACHERS' MARKS, MARKING SYSTEMS, ETC.

25. ASHBAUGH, E. J., "Reducing the Variability in Teachers' Marks," *Journal of Educational Research*, Vol. 9, pp. 185-198, 1924.

26. BANKER, H. J., "Significance of Teachers' Marks," *Journal of Educational Research*, Vol. 16, pp. 159-171 and 271-284, 1927.
An excellent discussion of marks and marking systems.

27. BLISS, D. C., "High School Failures," *Educational Administration and Supervision*, Vol. 3, pp. 125-138, 1917.

28. BOLTON, F. E., "Do Teachers' Marks Vary as Much as Supposed?" *Education*, Vol. 48, pp. 23-39, 1927. (T).
An attempted refutation of Starch's claim that teachers' marks vary enormously on the same paper.

29. BREEZE, R. E., "Correcting Examination Papers," *School Review*, Vol. 30, pp. 57-62, 1925.

30. CAMP, F. S., "Some 'Marks,' an Administrative Problem," *School Review*, Vol. 25, pp. 697-713, 1917.

31. CALDWELL, O. W. AND COURTIS, S. A., *Then and Now in Education 1845-1923*, World Book Co., 1924. 400 pp. (T).
Pages 37-46 give an interesting historical account of the views of Horace Mann on the written examination. Mann argues at length for the substitution of written for oral tests.

32. CATTELL, J. MCK., "Examinations, Grades, and Credits," *Popular Science Monthly*, Vol. 66, pp. 367-378, 1905.

33. COLVIN, S. S., "Marks and the Marking System as an Incentive to Study," *Education*, Vol. 32, pp. 560-572, 1912.

34. CORNING, G. B., "The Meaning of Students' Marks," *School Review*, Vol. 24, pp. 196-202, 1916.

35. DICKINSON, Z. C., "Suggestions Toward Improving Examination Marks," *University of Minnesota Bulletin*, Vol. 26, No. 31, pp. 31-36, 1923.

36. FEINGOLD, G. A., "Commutation of I Q's into Percentage Grades Corresponding to Those Commonly Used in Marking Scholarship," *Educational Administration and Supervision*, Vol. 11, pp. 251-263, 1925.
See also Symond's criticism, *loc. cit.*, pp. 264-266.

37. FINKELSTEIN, I. E., "The Marking System in Theory and Practice," *Educational Psychology Monographs*, No. 10, Warwick and York, 1913. 83 pp.
A study of the marking system at Cornell University, with a recommendation for a five-point plan with roughly fixed percentages.

38. FOSTER, W. T., "Scientific vs. Personal Distribution of College Credits," *Popular Science Monthly*, Vol. 78, pp. 388-408, 1911.

39. FRENCH, H. P., "A Practical Method of Translating Objective Scores into Percentage Marks," *Journal of Educational Method*, Vol. 6, pp. 60-61, 1926.
A ranking method of changing scores to percentage marks.

40. GAW, E. A., "College Grades," *School and Society*, Vol. 24, pp. 648-651, 1926.
Advocates sigma indices as marks in college classes.

41. GRAY, C. T., "Variations in the Grades of High-School Pupils," *Educational Psychology Monographs*, No. 8, 1913.
Plea for a scientific grading system as a remedy for the wide variations found in teachers' marks in ten public schools in Chicago and Indiana.

42. HENDRICKSON, C. E., "School Marks at Van Nuys High School," *Educational Research Bulletin*, Los Angeles City Schools, Vol. 7, No. 4, pp. 8-9, 1927. (T).

43. HULTEN, C. E., "The Personal Element in Teachers' Marks," *Journal of Educational Research*, Vol. 12, pp. 49-55, 1925.
Concludes that it is impracticable to attempt devices for equating teachers' marks, since a given teacher is not consistent in giving high or low marks.

44. JAMES, H. W., "A National Survey of the Grading of College Freshman Composition," *English Journal*, Vol. 15, pp. 579-587, 1926.

45. JAMES, H. W., "The Effect of Handwriting upon Grading," *English Journal*, Vol. 16, pp. 180-185, 1927.
Shows that poor handwriting lowered marks by seven points.

46. JOHNSON, F. W., "Comparative Study of Grades of Pupils from Different Elementary Schools in Subjects of the First-Year High School," *Elementary School Teacher*, Vol. 11, pp. 63-78, 1910.

47. JOHNSON, F. W., "A Study of High School Grades," *School Review*, Vol. 19, pp. 13-24, 1910.

48. KELLY, F. J., "Teachers' Marks, Their Variability and Standardization," Teachers College, Columbia University, *Contributions to Education*, No. 66, 1914. 139 pp. (T).
A widely quoted study of first importance, tending to prove the great variability and unreliability of marks.

49. LAUTERBACH, C. E., "Some Factors Affecting Teachers' Marks," *Journal of Educational Psychology*, Vol. 19, pp. 266-271, 1928.
Evidence is given that long-hand and typewritten identical papers show no significant differences in the marks received.

50. MERSEREAU, E. B., "A Study of the Significance of College Marks Considered as Ranks," *Educational Administration and Supervision*, Vol. 13, pp. 103-108, 1927.

51. MEYER, M., "The Grading of Students," *Science*, Vol. 28, pp. 243-252, 1908. (T).
An early but classic paper by the author of the so-called "Missouri system" of grading by means of the normal curve.

52. MILES, W. R., "Comparison of Elementary and High-School Grades," *University of Iowa Studies in Education*, Vol. 1. No. 1, 1911.

53. ODELL, C. W., "High-School Marking Systems," *School Review*, Vol. 33, pp. 346-354, 1925. (T).
Perhaps the best survey of marking practices. Shows the grading plans in use in 300 Illinois high schools, about 100 different systems being used.

54. RICH, S. G., "Economy and Fairness in Marking," *Journal of Educational Method*, Vol. 5, pp. 67-70, 1925.

55. RUCH, G. M. et al, *Objective Examination Methods in the Social Studies*, Scott, Foresman and Co., 1926. 116 pp.; especially pages 1-53. (T).
Experimental studies of official state eighth-grade examinations and the New York Regents' examinations, showing the influences of personal judgment in scoring, with resulting unreliability.

56. RUCH, G. M., *The Improvement of the Written Examination*, Scott, Foresman and Co., 1924. (T).
Pages 40-64 present a summary of the evidence on the unreliability of teachers' marks together with some evidence not published elsewhere.

57. RUGG, H. O., "Teachers' Marks and Marking Systems," *Educational Administration and Supervision*, Vol. 1, pp. 117-142, 1915.
Good summary up to this date.

58. SANDON, F., "A Statistical Analysis of Some School Marks," *Forum of Education*, Feb., 1925, pp. 24-31.

59. SHARP, L. A., "The Value of Standards in Grading Examinations," *Peabody Journal of Education*, Vol. 3, No. 1, pp. 38-45, 1925.
Adoption of a set of scoring rules reduced the variability of grading arithmetic papers by eighty per cent.

60. SHRINER, W. O., "The Reliability of Teachers' Marks," *Mathematics Teacher*, Nov. 19, 1924, pp. 426-443.

61. SPENCE, R. B., "The Improvement of College Marking Systems," Teachers College, Columbia University, *Contributions to Education*, No. 252, 1927. 89 pp.
An important monograph advocating a T-scale for marking.

62. STARCH, D. AND ELLIOTT, E. C., "The Reliability of Grading High-School Work in English," *School Review*, Vol. 20, pp. 442-457, 1912. (T).
This paper and the two following references constitute one of the most severe attacks on the reliability of teachers' marks. See Reference 28 (Bolton) for an attempted refutation.

63. STARCH, D. AND ELLIOTT, E. C., "The Reliability of Grading High-School Work in History," *School Review*, Vol. 21, pp. 676-681, 1913.

64. STARCH, D. AND ELLIOTT, E. C., "The Reliability of Grading High-School Work in Mathematics," *School Review*, Vol. 21, pp. 254-259, 1913. (T).

65. THORNDIKE, E. L., "Entrance Examinations and College Grades," *Science*, Vol. 23, pp. 839-845, 1906.
This paper and the following one are classic early demonstrations of the weaknesses in college entrance examinations.

66. THORNDIKE, E. L., "The Future of the College Entrance Examination Board," *Educational Review*, Vol. 31, pp. 470-479, 1906.

67. TIDYMAN, W. F., "Adjusting Marking Systems to Differences in Groups," *School and Society*, Vol. 22, pp. 247-248, 1925.

III. COMPARATIVE STUDIES OF OLD- AND NEW-TYPE TESTS

68. BARDY, J., "An Investigation of the Written Examination as a Measure of Achievement with Special Reference to General Science," University of Pennsylvania, 1923. 176 pp.

69. BATSON, W. H., "Reliability of the True-False Form of Examination," *Educational Administration and Supervision*, Vol. 10, pp. 95-103, 1924.

70. BLUMER, G., "Desirability of Changing the Type of Written Examinations," *Journal of the American Medical Association*, Vol. 72, pp. 1131-1133, 1919.

71. BOYD, W., "Exploration of the True-False Method of Examination," *Forum of Education*, Vol. 4, pp. 34-38, 1926.

72. BRINKLEY, S. G., "Values of the New-Type Examinations with Special Reference to History." (See Reference 2.)

73. CRAWFORD, C. C. AND RAYNALDO, D. A., "Some Experimental Comparisons of True-False Tests and Traditional Examinations," *School Review*, Vol. 33, pp. 698-706, 1925. (T).
In twenty comparisons, fifteen favored the older examination over the true-false.

74. FISKE, T. S., "Annual Reports of the Secretary of the College Entrance Examination Board," 1921-1924. N. Y., 431 West 117th St.
Several papers giving the evidence leading to the adoption, in part, of objective examinations by this body.

75. GATES, A. I., "The True-False Test as a Measure of Achievement in College Courses," *Journal of Educational Psychology*, Vol. 12, pp. 267-287, 1921.
Presents evidence that reliability of true-false tests exceeds that of the essay examination.

76. GLENN, E. R., "The Conventional Examination in Chemistry and Physics Versus the New Type of Tests," *School Science and Mathematics*, Vol. 21, pp. 666-670 and 746-756, 1921; and Vol. 23, pp. 459-470, 1923.

77. HANNIG, W. A., "Relative Worth of Short Answer and Free Answer Material in Elementary Teacher Tests," *Public Personnel Studies*, Vol. 4, pp. 277-278, 1926.

78. JAMES, B. J., "The Modern Test," *School and Society*, Vol. 23, pp. 209-213, 1924.
Proposes frequent new-type tests instead of lengthy essay examinations at end of year, and gives reasons for this recommendation.

79. KNIGHT, F. B., "Data on the True-False Test as a Device for College Examinations," *Journal of Educational Psychology*, Vol. 13, pp. 75-80, 1922.

80. LAIRD, D. G., "A Comparison of the Essay and the Objective Type of Examinations," *Journal of Educational Psychology*, Vol. 14, pp. 123-124, 1923.

81. LOHR, V. C., "A Comparison of Some Tests Given in High-School Physics," *School Science and Mathematics*, Vol. 27, pp. 74-85, 1927.

82. PATERSON, D. G., "Do New and Old Examinations Measure Different Functions?" *School and Society*, Vol. 24, pp. 246-248, 1926.

83. ROBACK, A. A., "Subjective Tests vs. Objective Tests," *Journal of Educational Psychology*, Vol. 12, pp. 439-444, 1921.

84. SANFORD, V., "A New-Type Final Geometry Examination," *The Mathematics Teacher*, Vol. 18, pp. 23-36, 1925.

85. SCHRYOCH, R. H., "New Tests for Old," *Historical Outlook*, Vol. 14, pp. 319-322, 1923.

86. SKINNER, A. W., "Examinations and Tests," *High School Quarterly*, Vol. 14, pp. 174-179, 1926.
Concludes that old and new types have their special advantages and that both are needed.

87. THARP, J. B., "The New Examination Versus the Old in Foreign Language," *School and Society*, Vol. 26, pp. 691-694, 1927.

88. WOOD, B. D., "The New Type Examinations in the College of Physicians and Surgeons," *Journal of Personnel Research*, Oct., Nov., 1926, pp. 227-234 and 277-283.
An important experimental paper.

89. WOOD, B. D., "The Measurement of Law School Work," *Columbia Law Review*, Vol. 24, No. 3, March, 1924; and Vol. 25, No. 3, March, 1925.
An account of the application of objective tests in law courses, together with the opinions of the instructors in charge.

IV. INSTRUCTIONAL USES OF OBJECTIVE TESTS

90. BUTLER, W. F., "The Value of Informal Tests in Supervision," *First Yearbook, Elementary School Principals*, 1922, pp. 94-119.

91. ELSTON, B., "Improving the Teaching of History Through the Use of Tests," *Historical Outlook*, Vol. 14, pp. 300-305, 1923.

92. GRAY, W. S., "Value of Informal Tests of Reading Accomplishment," *Journal of Educational Research*, Vol. 1, pp. 103-111, 1920.
Advocates informal tests constructed by the teacher, and holds that standard tests alone cannot fill such needs.

93. HEFFERNAN, H., "Objective Measurement of Progress in Country Schools," M. A. Thesis, University of California, 1925.

94. HERRING, J. P., "Educative Control by Means of a New-Type Measurement," *Journal of Educational Method*, Vol. 4, pp. 94-102, 1924.

95. KIMMEL, W. G., "Testing Pupil Progress in Community Life English," *Supplementary Educational Monographs*, 26, pp. 33-69, 1925.

96. MCLEOD, B. AND IRVING, H., "Objective Examinations in the Rural Schools of Wyoming," *Journal of Educational Research*, Vol. 17, pp. 45-49, 1928.

97. MALONEY, E. AND RUCH, G. M., "The Use of Objective Tests in Teaching as Illustrated by Grammar," *The School Review*, Vol. 37, pp. 62-66, 1929.

98. SCHUTTE, T. H., "Is There Value in the Final Examination?" *Journal of Educational Research*, Vol. 12, pp. 204-213, 1925.
A group given examinations showed greater accomplishment than a similar group taking no examinations.

99. SPENCER, P. L., "The Improvement of Teaching by Means of 'Home-Made' Non-standardized Diagnostic Tests and Remedial Instruction," *School Review*, Vol. 31, pp. 276-281, 1923.
A very thoughtful discussion of diagnostic and remedial teaching through well-made tests.

100. SUTHERLAND, A. H., "Tests in Reading as a Part of Classroom Routine," *Nineteenth Yearbook of the National Society for the Study of Education*, Part I, pp. 47-51, 1920.

101. WAPLES, D., "The Best-Answer Exercise as a Teaching Device," *Journal of Educational Research*, Vol. 15, pp. 10-22, 1927.
Suggests the best answer test as a measure of the organization of data, and gives a number of illustrations.

102. WOODY, C. W., "Informal Tests as a Means for the Improvement of Instruction," *First Yearbook, Department of Elementary School Principals*, N. E. A., pp. 87-94, 1922.
A discussion which will well repay reading.

V. STUDENTS' ATTITUDES TOWARD EXAMINATIONS

103. BARDY, J. (See Reference 68.) (T).

104. BRINKLEY, S. G. (See Reference 2). (T).

105. BUCKNER, C. A. AND HUGHES, R. O., "Testing Results in the Social Studies," *Journal of the School of Education*, University of Pittsburgh, Vol. 1, No. 1, pp. 5-11, 1925. (T).

106. KINDER, J. S., "Supplementing Our Examinations," *Education*, Vol. 45, pp. 557-566, 1925. (T).

107. KLISE, N. M., "Student Opinion of Type of Examination," *School and Society*, Vol. 24, pp. 23-24, 1926.
The preferences of 337 students and their reasons are given.

108. KOLSTOE, S. O., "Reactions to True-False Tests," *School of Education Record*, University of North Dakota, 11, pp. 54-55, 1926. (T).

109. MAY, M. A., "Measuring Achievement in Elementary Psychology and in Other College Subjects," *School and Society*, Vol. 17, pp. 472-476 and 556-560, 1923. (T).

110. OZANNE, C. E., "A Study of Different Types of Teachers' Tests," *School Review*, Vol. 34, pp. 54-60, 1926.
A study of the relative interests in old and new types of examinations.

111. SOMERS, G. T., "Students' Attitudes Toward Examinations," *Bulletin of the School of Education*, Indiana University, Vol. 3, No. 1, pp. 1-48, 1926. (T).

VI. SAMPLES OF OBJECTIVE TESTS[1]

112. ABBOTT, A., "Tests for English Teachers," *The English Journal*, Vol. 12, pp. 663-671, 1923.

113. BERG, G., "Construction of Tests on *Silas Marner*," *University High School Journal*, Oakland, California, Vol. 7, pp. 304-317, 1928.
A very good set of test items on *Silas Marner*.

114. BERNSTEIN, L., "A New-Type Examination in History," *New York Bulletin of High Points*, March, 1923, pp. 15-20.

115. BRIGGS, T. H., "A Dictionary Test," *Teachers College Record*, Vol. 24, pp. 355-365, 1923.
Good suggestions for constructing tests on the use of the dictionary.

[1]See also the references listed in Sections I and IX of this bibliography.

116. BRIGGS, T. H., "An Examination in First-Year Latin," *The Classical Weekly*, Vol. 16, No. 19, 1923.

117-157. The following references present suggested tests, unstandardized or partially standardized, prepared by the Bureau of Public Personnel Administration, Washington, D. C., under the direction of Mr. Fred Telford. All citations refer to volumes of *Public Personnel Studies*. Exact titles of articles are not given, but the nature of the test is stated in each instance.

Vol. II, 1924, No. 3	Tests for prison guard
Vol. II, 1924, No. 4	Tests for patrolman
Vol. II, 1924, No. 5	Tests for hospital attendant
Vol. II, 1924, No. 6	Tests for senior clerk
Vol. II, 1924, No. 7	Tests for fire fighter
Vol. II, 1924, No. 8	Tests for food inspector
Vol. II, 1924, No. 9	Tests for supervising clerk
Vol. III, 1925, No. 1	Tests for janitor
Vol. III, 1925, No. 2	Test in bacteriology
Vol. III, 1925, No. 3	Tests for senior library assistant, circulation department
Vol. III, 1925, No. 4	Tests for painter
Vol. III, 1925, No. 5	Tests for automobile driver
Vol. III, 1925, No. 6	Tests in pathology
Vol. III, 1925, No. 8	Tests for female playground supervisor
Vol. III, 1925, No. 9	Tests for road inspector
Vol. III, 1925, No. 10	Tests for plumber
Vol. III, 1925, No. 11	Tests for shift engineman
Vol. III, 1925, No. 12	Tests for junior clerk
Vol. IV, 1926, No. 2	Tests for private branch exchange operator
Vol. IV, 1926, No. 3	Tests for electrician
Vol. IV, 1926, No. 4	Tests for patrolman
Vol. IV. 1926, No. 5	Tests for senior account clerk
Vol. IV, 1926, No. 6	Tests for probation officer
Vol. IV, 1926, No. 7	Tests for general machinist
Vol. IV, 1926, No. 8	Tests for automobile mechanic
Vol. IV, 1926, No. 10	Tests for elementary teacher
Vol. IV, 1926, No. 11	Tests for carpenter
Vol. IV, 1926, No. 12	Tests for vegetable gardener
Vol. V, 1927, No. 1	Tests for fire lieutenant
Vol. V, 1927, No. 2	Tests for instrument man
Vol. V, 1927, No. 3	Tests for police sergeant
Vol. V, 1927, No. 4	Tests of alphabetical filing and ability to follow directions
Vol. V, 1927, No. 5	Tests for steam firemen
Vol. V, 1927, No. 6	Tests for cook
Vol. V, 1927, No. 7	Tests for senior cook
Vol. V, 1927, No. 10	Tests for junior personnel examiner

Vol. V, 1927, No. 11 Tests for laboratory assistant
Vol. V, 1927, No. 12 Tests for selection of policewoman
Vol. VI, 1928, No. 1 Revised tests for food inspector
Vol. VI, 1928, No. 2 Preliminary tests for stenographer
Vol. VI, 1928, No. 4 Tests for blacksmith

158. CARLSON, P. A., "A Test Program in Bookkeeping," *The Balance Sheet*, Vol. 7, No. 1, pp. 12-14, 1925.

159. CARLSON, P. A., "Tests and Measurements in Bookkeeping," *Bulletin of the Whitewater State Normal School*, Whitewater, Wisconsin, Vol. 11, No. 1, Bulletin No. 119, pp. 3-9, 1925.
The Southwestern Publishing Company of Cincinnati, Ohio, publishes a number of commercial tests devised by Mr. P. A. Carlson.

160. CHAPMAN, J. C. and TOOPS, H. A., "A Written Trade Test; Multiple-Choice Method," *Journal of Applied Psychology*, Vol. 3, pp. 358-365, 1919.
One of the earliest articles on the new-type examination.

161. CHAPMAN, J. C., "The Measurement of Physics Information," *School Review*, Vol. 27, pp. 748-756, 1919.

162. CHASSELL, C. F. AND E. B., "A Test and Teaching Device in Citizenship for Use in Junior High Schools," *Educational Administration and Supervision*, Vol. 10, pp. 7-29, 1924.

163. COOLEY, A. M. AND REEVES, G., "Some Investigations Concerning the Use of Certain Home Economics Information Tests," *Teachers College Record*, Vol. 24, pp. 374-392, 1923.

164. COOPRIDER, J. L., "Exercises in Biology," *School Science and Mathematics*, Vol. 25, pp. 807-813, 1925.

165. DALMAN, M. A., "Hurdles, a Series of Calibrated Objective Tests in First-Year Algebra," *Journal of Educational Research*, Vol. 1, pp. 47-62, 1920.

166. DUBREUIL, A. J., "True-False Tests in Literature and Formal English," *Bulletin of the Illinois Association of Teachers of English*, Vol. 15, pp. 1-17, 1923.

167. EATON, M. P., "New Style Examinations in English at Wadleigh High School," *New York Bulletin of High Points*, Vol. 5, No. 4, pp. 3-16, 1923.

168. EVANS, M. E., "Objective Tests in Eighth-Grade Literature," *Elementary English Review*, Vol. 5, pp. 13-22, 1928.

169. FARWELL, H. W., "The New-Type Examination in Physics," *School and Society*, Vol. 19, pp. 315-322, 1924.

170. FILER, H. A. AND O'ROURKE, L. J., *Annual Reports of the Chief Examiner and Director of Research of the U. S. Civil Service Commission for the Fiscal Year Ended June 30, 1923.* Washington, 1923, Government Printing Office.
A description of objective examination methods in the Civil Service.

171. GATES, A. I. AND STRANG, R., "A Test in Health Knowledge," *Teachers College Record*, Vol. 26, pp. 867-880, 1925.

172. GOODALE, M., "New-Type Tests," *Journal of Geography*, Vol. 27, pp. 63-70, 1928.

173. GORDON, H. C., "Some New-Type Forms in High-School Physics," *School Science and Mathematics*, Vol. 27, pp. 721-733, 1927.

174. "Kansas Scholarship Contest, Test Questions for 1923," *Teaching*, Emporia, Kansas, Vol. 7, No. 65, pp. 13-39, 1923 and Vol. 7, No. 67, pp. 1-69, 1924.

175. KARWOSKI, T. R. AND CHRISTENSEN, E. O., "A Test of Art Appreciation," *Journal of Educational Psychology*, Vol. 17, pp. 187-194, 1926.

176. LATHROP, H. O., "Testing in Commercial Geography," *Journal of Geography*, Vol. 26, pp. 256-262, 1927.

177. LAYCOCK, S. R., "The Laycock Test of Biblical Information," *Journal of Educational Psychology*, Vol. 16, pp. 329-334, 1925.

178. LOHR, V. C., "A Comparison of Some Tests Given High-School Students in Physics," *School Science and Mathematics*, Vol. 27, pp. 74-85, 1927.

179. MAY, M. A., "Standardized Examinations in Psychology and Logic," *School and Society*, Vol. 11, pp. 553-540, 1920.

180 MCCLUSKY AND DOLCH, E. W., "A Study Outline Test," *School Review*, Vol. 32, pp. 757-773, 1924.
A test of the ability of the student to read and comprehend the thought of an author.

181. MELVIN, A. G., "A True-False Test in English Literature," *The English Journal*, Vol. 11, pp. 491-496, 1922.

182. MILLER, W. S., "An Objective Test in Educational Psychology," *Journal of Educational Psychology*, Vol. 16, pp. 237-246, 1925.

183. MOYER, F. E., "New Types of History Tests," *Historical Outlook*, Vol. 14, pp. 323-324, 1923.

184. ODELL, C. W., "Objective Measurement of Information," *University of Illinois Bulletin*, Vol. 23, No. 36, *Bureau of Educational Research Circular*, No. 44, 1926, 27 pp.
An important discussion of 37 types of objective tests, with illustrations.

185. ORLEANS, J. S., "Manual on the Local Construction and Uses of Objective Tests," *University of the State of New York Bulletin*, No. 893, Feb. 1, 1928, 57 pp.
A great many examples of objective tests are given together with valuable general discussions.

186. SANFORD, V., "A New-Type Final Geometry Examination," *Mathematics Teacher*, Vol. 18, pp. 22-37, 1925.

187. STACK, H. J., "Standardized Tests in Community and Economic Civics," *Historical Outlook*, Vol. 18, pp. 166-172, 1927.

188. STRICKLAND, V. L., "Objective Tests," *Kansas State Agricultural College Bulletin*, Vol. VIII, No. 2, 23 pp.

189. THORNDIKE, E. L., "Completion Tests in Physics," *School Science and Mathematics*, Vol. 22, pp. 637-647, 1922.

190. TOOPS, H. A., "A General Science Test," *School Science and Mathematics*, Vol. 25, pp. 817-822, 1925.

191. VARIOUS, "Objective Test Number," *The High School*, School of Education, University of Oregon, Eugene, Oregon, Vol. 4, No. 4, pp. 119-158, 1927.
An excellent collection of objective tests for almost all high-school subjects.

192. VARIOUS, "New-Type Examination Questions" (Library Methods), Summer Library Institute, University of Chicago, 1926. (T).

193. VARIOUS, "Tests in History and the Social Studies," *Historical Outlook*, November, 1923.
Practically the entire issue of this journal is devoted to examples of objective tests.

194. WEBB, H. A., "Testing Laboratory Resourcefulness," *School Science and Mathematics*, Vol. 22, pp. 259-267, 1922.

195. WRIGHT, H. C., "Some True-False Examinations for Use in General Mathematics," *Mathematics Teacher*, Vol. 28, pp. 83-91, 1925.

VII. EXPERIMENTAL AND THEORETICAL PAPERS

196. ARNOLD, H. L., "Analysis of Discrepancies Between True-False and Simple Recall Examinations," *Journal of Educational Psychology*, Vol. 18, pp. 414-420, 1927.

197. ASKER, W., "The Reliability of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 9, pp. 234-240, 1924. (T).

198. BALLARD, P. B. (See Reference 1, especially pp. 96-98). (T).

199. BARTHELMESS, H. M., "Reply to a Criticism of Tests Requiring Alternative Reponses," *Journal of Educational Research*, Vol. 6, pp. 357-359, 1922. (T).

200. BORING, E. G., "The Logic of the Normal Law of Error in Mental Measuring," *American Journal of Psychology*, Vol. 31, pp. 1-33, 1920.

201. BRINKLEY, S. G., "Relative Value of Different Types of Questions in Reading Tests," *School Science and Mathematics*, Vol. 25, pp. 703-708, 1925.

202. BRINKLEY, S. G. (See Reference 2, especially pp. 64-82 and 84-90). (T).

203. BURTON, W. H., "A Contribution to the Technique of Constructing 'Best-Answer' Tests," *Elementary School Journal*, Vol. 22, pp. 762-770, 1925.

204. CHAPMAN, J. C., "Individual Injustice and Guessing in the True-False Examination," *Journal of Applied Psychology*, Vol. 6, pp. 342-348, 1922.

205. CHRISTENSEN, A. M., "A Suggestion as to Correcting Guessing in Examinations," *Journal of Educational Research*, Vol. 14, pp. 370-374, 1926. (T).

206. CRAWFORD, C. C. AND RAYNALDO, D. A. (See Reference 73.) (T).

207. DOUGLASS, H. R. AND SPENCER, P. L., "Is it Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology*, Vol. 14, pp. 109-112, 1923. (T).

208. DUBREUIL, A. J., "An Answer to Professor Leonard's Dangers of the True-False Test in English," *Bulletin of the Illinois Association of Teachers of English*, Vol. 16, No. 7, pp. 1-5, 1924.

209. FORAN, T, G., "Methods of Scoring Alternative-Response and Multiple-Choice Tests," *Catholic Educational Review*, February, 1925, pp. 1-8.

210. FOSTER, R. R. AND RUCH, G. M., "On Corrections for Chance in Multiple-Response Tests," *Journal of Educational Psychology*, Vol. 18, pp. 48-51, 1927. (T).

210. FRITZ, M. F., "Guessing in a True-False Test," *Journal of Educational Psychology*, Vol. 18, pp. 558-561, 1927. (T).
The author found that difficult true-false items on medical terms, unknown to the students, were answered true and false in the ratio 62:38. The same ratio held for familiar materials.

211. GATES, A. I., "The True-False Test as a Measure of Achievement," *Journal of Educational Research*, Vol. 12, pp. 276-287, 1921.

212. GILES, J. T., "Improving the Objective Test Question," *School Review*, Vol. 35, pp. 286-288, 1927.
Advocates elimination of poorest response rather than the selection of the best in multiple-choice tests.

213. GREENE, H. A., "A New Correction for Chance in Examinations of the Alternate-Response Type," *Journal of Educational Research*, Vol. 17, pp. 102-107, 1928. (T).

214. GRIFFIN, H. D., "Safeguarding the Final Examination," *School and Society*, Vol. 23, pp. 343-344, 1926.
The author recommends presenting the test pages bound in varying orders to prevent cheating.

215. GUNDLACH, R., "A Method for the Detection of Cheating in College Examinations," *School and Society*, Vol. 22, pp. 215-216, 1925.
It is recommended that different sets of the same questions be presented with the order of items varying.

216. HAHN, H. H., "A Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 6, pp. 236-240, 1922. (T).

217. HAMMOND, E. L., "A Study of the Reliability of an Objective Examination in Ninth-Grade English," *School Review*, Vol. 35, pp. 45-51, 1927.

220. HOLZINGER, K. J., "On Scoring Multiple-Response Tests," *Journal of Educational Psychology*, Vol. 15, pp. 445-447, 1924. (T).

221. JAMES, H. W., "Technique in the Construction of a Teacher's Own Objective Tests," *Peabody Journal of Education*, Vol. 4, pp. 240-243, 1927.

222. KOHS, S. C., "High Test Scores Obtained by Subaverage Minds," *Psychological Bulletin*, Vol. 17, pp. 1-5, 1920. (T).

223. LAIRD, D. A., "A Note on the Shortening of the Examination," *Journal of Educational Psychology*, Vol. 15, pp. 116-117, 1924.

224. LEE, B., "Some Faults Common to Informal Objective Tests Made by High-School Teachers," *Educational Administration and Supervision*, Vol. 14, pp. 105-113, 1928.
A number of faults were found and listed. Many of these are obvious and most have been previously reported by Weidemann and others. These precautions are nevertheless very important.

225. LEHMANN, H. C., "Does it Pay to Change Initial Decisions in a True-False Test?" *School and Society*, Vol. 28, pp. 456-458, 1928.
Superior students tended to better their scores by revision of answers, the reverse being true for inferior students. In general, the larger the number of changes the less likelihood of betterment of marks.

226. MCAFEE, L. O., "The Reliability of Non-Standardized Point Scales," *Elementary School Journal*, Vol. 24, pp. 579-585, 1924.

227. MCCALL, W. A., *How to Measure in Education*, 1922, The Macmillan Company; especially pp. 121-126.

228. MCCLUSKY, H. Y. AND CURTIS, F. D., "A Modified Form of the True-False Test," *Journal of Educational Research*, Vol. 14, pp. 213-225, 1926. (T).

229. MATHEWS, C. O.,"The Effect of Position of Printed Response Words upon Children's Answers in Two-Response Types of Tests," *Journal of Educational Psychology*, Vol. 18, pp. 445-457, 1927. (T).

230. MAY, M. A., "Measuring Achievement in Elementary Psychology and Other College Subjects," *School and Society*, Vol. 17, pp. 472-476 and 556-560, 1923. (T).

231. MILLER, G. F., "A Variation in the True-False Achievement Test," *School and Society*, Vol. 20, pp. 250-251, 1924.

232. MILLER, G. F., "Formulas for Scoring Tests in Which the Maximum Amount of Chance is Determined," *Journal of Educational Psychology*, Vol. 16, pp. 304-315, 1925.
A contribution to the methods of correcting for chance.

233. MONROE, W. S., "Written Examinations Versus Standardized Tests," *School Review*, Vol. 34, pp. 253-265, 1924.

234. MUENZINGER, K. F., "Critical Note on the Reliability of a Test," *Journal of Educational Psychology*, Vol. 18, pp. 424-428, 1927.
A statistical discussion of the concept of test reliability.

235. ODELL, C. W., "Another Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 8, pp. 326-330, 1923. (T).

236. PATERSON, D. G. AND LANGLIE, T. A., "Empirical Data on the Scoring of True-False Tests," *Journal of Applied Psychology*, Vol. 9, pp. 339-348, 1925. (T).

237. REMMERS, H. H. *et al*, "An Experimental Study of the Relative Difficulty of True-False, Multiple-Choice, and Incomplete-Sentence Types of Examination Questions," *Journal of Educational Psychology*, Vol. 14, pp. 366-372, 1923. (T).

238. REMMERS, H. H. AND E. M., "The Negative Suggestion Effect of True-False Examination Questions," *Journal of Educational Psychology*, Vol. 17, pp. 52-56, 1926. (T).

239. RICH, G. J., "A Scale for Scoring Tests with Alternative Responses," *American Journal of Psychology*, Vol. 36, pp. 597-600, 1925.

240. RICHARDS, O. W., "High Test Scores Attained by Subaverage Minds," *Journal of Experimental Psychology*, Vol. 7, pp. 148-156, 1924. (T).

241. RICHARDS, O. W. AND KOHS, S. C., "High Test Scores Attained by Subaverage Minds," *Journal of Educational Psychology*, Vol. 16, pp. 8-17, 1925. (T).

242. ROBERTS, H. M. AND RUCH, G. M., "The Negative Suggestion Effect of True-False Tests," *Journal of Educational Research*, Vol. 18, pp. 112-116, 1928. (T).

243. RUCH, G. M., *The Improvement of the Written Examination*, especially pp. 114-121. (See Reference 16.) (T).

244. RUCH, G. M. AND CHARLES, J. W., "A Comparison of Five Types of Objective Tests in Elementary Psychology," *Journal of Applied Psychology*, Vol. 12, pp. 398-403, 1928. (T).

245. RUCH, G. M. AND DEGRAFF, M. H., "Corrections for Chance and 'Guess' vs 'Do Not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology*, Vol. 17, pp. 368-375, 1926. (T).

246. RUCH, G. M., et al, *Objective Examination Methods in the Social Studies*, pp. 54-88. (See Reference 17.) (T).

247. RUCH, G. M. *et al*, "Short-Answer Examinations in the Social Studies in the Elementary School Grades," *Public Personnel Studies*, Vol. 4, No. 10, pp. 274-277, 1926.

248. RUCH, G. M. AND STODDARD, G. D., "Comparative Reliabilities of Five Types of Objective Examinations," *Journal of Educational Psychology*, Vol. 16, pp. 89-103, 1925. (T).

249. RUCH, G. M. AND STODDARD, G. D., *Tests and Measurements in High School Instruction*, pp. 282-294. (See Reference 15.) (T).

250. RUTLEDGE, R. E., "The True-False Examination in Elementary Psychology with Suggestions for its Improvement," Ph. D. Thesis, University of California, 1926. (T).

251. SYMONDS, P. M., "Factors Influencing Test Reliability." *Journal of Educational Psychology*, Vol. 19, pp. 73-87, 1928.

Perhaps the best discussion of the concept of test reliability now in print, especially for non-statistical readers.

252. TELFORD, F. (Unsigned), "The So-called Guessing Element in Written Tests," *Public Personnel Studies*, Vol. 3, No. 10, pp. 275-282.

253. THURSTONE, L. L., "A Method of Scoring Tests," *Psychological Bulletin*, Vol. 16, pp. 235-240, 1919. (T).

254. TOOPS, H. A., "How to Construct an Objective Test," *School Science and Mathematics*, Vol. 25, pp. 817-822, 1925.

255. TOOPS, H. A., "Trade Tests in Education." (See Reference 21.) (T).

256. U. S. WAR DEPARTMENT, ADJUTANT GENERAL'S OFFICE, "The Making of a Trade Test," *Personnel Bulletin*, Vol. 1, No. 7, pp. 25-28, 1918.

257. Weidemann, C. C., "Determinate vs. Indeterminate Information in Written Examinations," *Journal of Educational Method*, Vol. 7, pp. 126-127, 1927.

258. WEIDEMANN, C. C., "Limitations of the True-False Statement," *Journal of Educational Method*, Vol. 7, pp. 214-215, 1928.

259. WEINLAND, J. D., "A Note on the Right-Wrong Examination," *Journal of Educational Psychology*, Vol. 18, pp. 266-268, 1927.

260. WEST, P. V., "A Critical Study of the Right-Minus-Wrong Method," *Journal of Educational Research*, Vol. 8, pp. 1-9, 1923. (T).

261. WEST, P. V., "The Significance of Weighted Scores," *Journal of Educational Research*, Vol. 15, pp. 302-308, 1924.
Like Douglass and Spencer, West finds weighting of little value.

262. WIGMORE, J. H., "The 'New-Type' Law Examination," *Illinois Law Review*, Vol. 19, pp. 172-173, 1923.

263. WILSON, H. E., "The Continuity Test in History Teaching," *School Review*, Vol. 34, pp. 679-684, 1926.

264. WOOD, B. D., *New York Experiments with New-Type Modern Language Tests*. (See Reference 24.) (T).

265. WOOD, B. D., "Studies of Achievement Tests," *Journal of Educational Psychology*, Vol. 17, pp. 1-22, 125-139, and 263-269, 1926. (T).

266. WOOD, B. D., "The Measurement of College Work," *Educational Administration and Supervision*, Vol. 7, pp. 301-331, 1921. (T).

267. WOOD, E. P., "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Pscychology*, Vol. 18, pp. 18-25, 1927. (T).

268. WORCESTER, D. A., "Prevalent Errors in New-Type Examinations," *Journal of Educational Research*, Vol. 18, pp. 48-52, 1928.
A great many pitfalls are listed with concrete examples. A very important discussion for the beginner at objective testing.

269. YERKES, R. M. (Ed.), "Psychological Examining in the U. S. Army," *Memoirs of the National Academy of Sciences*, Vol. 15, pp. 305 and 339, 1921. Washington, Government Printing Office. (T).

VIII. Miscellaneous[1]

270. Achtenhagen, O., "Why is an Examination—and What of It?", *English Journal*, Vol. 15, pp. 285-289, 1926.

271. Bawden, W. T., "The Army Trade Tests," *U. S. Bureau of Education Circular*, No. 4, April, 1919. 28 pp.

272. Bingham, W. V., "Measuring a Workman's Skill; the Use of Trade Tests in Army and Industrial Establishments," *Proceedings of the National Society for Vocational Education*, Bulletin 30, 1919.

273. Bird, C., "The Detection of Cheating in Objective Examinations," *School and Society*, Vol. 25, pp. 261-262, 1927.

274. Blumer, G., "Desirability of Changing the Type of Written Examinations," *Journal of the American Medical Association*, Vol. 62, pp. 1131-1133, 1919.

275. Branom, M. E., *The Measurement of Achievement in Geography*, N. Y., The Macmillan Company, 1925. 186 pp.

276. Brown, C. M., "Construction and Use of Information Tests in Home Economics,", *Journal of Home Economics*, Vol. 16, pp. 251-256, 1924.

277. Brown, C. M., "What Can Educational Measurements Do for Home Economics?", *Journal of Home Economics*, Vol. 16, pp. 191-196, 1924.

278. Buckner, C. A. and Hughes, R. O., "Testing Results in Social Studies," University of Pittsburgh, *School of Education Journal*, Vol. 1, No. 1, pp. 5-12, 1925.

279. Caldwell, O. W. and Courtis, S. A., *Then and Now in Education, 1845-1923*, N. Y., The World Book Company, 1924. (T).

280. Charters, W. W., "Constructing a Language and Grammar Scale," *Journal of Educational Research*, Vol. 1, pp. 249-258, 1920.
An account of the validation of language tests of the standardized variety.

281. Cheydleur, F. D., "Construction and Validation of a French Grammar Test of the Selection and Multiple-Choice Type," *Journal of Educational Research*, Vol. 17, pp. 184-196, 1928.
Another good account of the validation of an objective test.

282. College Entrance Examination Board, *Report of the Commission on New Type Examinations*, April, 1924.
A very significant paper in the light of the history of the type of examinations set by the College Entrance Examination Board.

283. Dadourian, H. M., "Are Examinations Worth the Price?", *School and Society*, Vol. 21, pp. 442-443, 1925.

284. Dickinson, Z. C., "Suggestions Toward Improving Examination Marks," *Bulletin of the University of Minnesota*, Vol. 26, No. 31, pp. 31-36, August 4, 1923.

[1]Many of the articles listed in Section VIII might have been classified elsewhere. A number of articles appearing after Sections I to VII were prepared have been included here.

285. DOUGLASS, H. R., "Quizzes, Examinations, and Marking," *Modern Methods in High School Teaching*, Boston, Houghton Mifflin Company, pp. 357-391, 1926. *Ibid.*, "New Ideas in Written Examinations," pp. 427-456.

286. DOYLE, L. AND FOOTE, M., "The Pledge as an Instrument to Secure Honesty in Examinations," *Peabody Journal of Education*, Vol. 3, pp. 79-84, 1925.
Results gathered in one school indicate the undesirability of requiring the pledge of honesty.

287. ELLIS, R. S., "A Method of Constructing and Scoring Tests Given with Time Limits to Eliminate or Weight the Effect of Speed," *School and Society*, Vol. 28, pp. 205-207, 1928.

288. FENTON, N. AND LEHMAN, H. C., "The True-False Question and the Student's Sense of Honesty," *School and Society*, Vol. 28, pp. 115-116, 1928.
Present a plan for allowing students to qualify answers on true-false tests.

289. FLACK, W. S., "Investigation of Mathematical Ability in the Classroom," *Forum of Education*, Vol. 4, pp. 44-56, 1926.

290. GATHANY, J. M., "The Giving of History Examinations," *Education*, Vol. 34, pp. 514-521, 1914.

291. GERRY, H. L., "Types of Tests Desirable for Chemistry and the Present Status of Their Development," *School Science and Mathematics*, Vol. 25, pp. 918-922, 1925.

292. GLENN, E. R., "The Conventional Examination in Chemistry and Physics vs. the New-Type Tests," *School Science and Mathematics*, Vol. 21, pp. 671-675 and 746-756, 1921; Vol. 23, pp. 459-470, 1923.

293. GOOD, H. G., "A Form of Matching Test," Ohio State University, *Educational Research Bulletin*, Vol. 6, No. 8, pp. 158-160, 1927.
A variate of the matching test in which pairs of items, when properly matched, stand in relation of subject and predicate of a sentence.

294. HANCE, R. T., "Mental Ability and the Examination," *School and Society*, Vol. 20, pp. 445-446, 1924.

295. HATCH, R. W., "How Can Examinations in History Be Improved?", *Horace Mann Studies in Education*, Columbia University, New York City, February, 1922.

296. HAWKES, H., "Experimenting with the New Type of Examination at Columbia," *School and Society*, Vol. 15, pp. 141-142, 1922.

297. HEMSTREET, A. E., "Justice in Testing," *The School Magazine*, Vol. 8, No. 5, pp. 220-221, 1926.

298. INGLIS, A., "Variability in Judgments in Equating Values in Grading," *Educational Administration and Supervision*, Vol. 2, pp. 25-30, 1916.

299. JENNISON, H. M., "Improvement in Examination Technique for Teachers of Botany," *School Science and Mathematics*, Vol. 27, pp. 832-843 and 944-951, 1927.

300. JOHNSON, F. W., "The Marking System," *The Administration and Supervision of the High School*, Boston, Ginn and Company, Chapter 15, 1925.

301. KEPNER, T. P., "A Survey of the Test Movement in History," *Journal of Educational Research*, Vol. 7, pp. 309-325, 1923.

302. KIMBALL, E. P., "As for Examinations," *School and Society*, Vol. 27, pp. 571-572, 1928.

303. KIMMEL, W. C., "Practice Tests in the Social Studies," *Historical Outlook*, Vol. 14, pp. 354-358, 1923.

304. LOWELL, A. L., "The Art of Examination," *Atlantic Monthly*, Vol. 137, pp. 58-66, 1926.

305. MCCALL, W. A., "A New Kind of School Examination," *Journal of Educational Research*, Vol. 1, pp. 34-46, 1920.

306. MACPHAIL, A. H., "Measuring Achievement in High School," *American Educational Digest*, Vol. 46, pp. 357-359 and 378, 1927.

307. MARSH, W. R., et al, *Report of the Commission on New Types of Examinations*, etc., New York, The College Entrance Examination Board, 1923. 39 pp.

308. MEAD, A. R., "Suggestions for the Training of Teachers in the Use of Educational Measurements," *Educational Administration and Supervision*, Vol. 12, pp. 23-43, 1926.

309. MORLEY, E. E., "Final Examinations and the Effect of Exemptions," *High School Teacher*, Vol. 2, pp. 90-91, 1926.

310. MOYER, F. E. *et al*, "Middle States Conference on History Tests," *The Historical Outlook*, Vol. 14, pp. 323-328, 1923.
Advantages and economy of objective tests are discussed.

311. OPDYKE, J. B., "Constructive Examinations," *Educational Review*, Vol. 73, pp. 33-43, 1927.

312. OGAN, R. W. AND EWING, D. H., "Reducing Variability in Teachers' Scoring," Ohio State University, *Educational Research Bulletin*, Vol. 7, No. 10, pp. 214-216, 1928.
Shows the value of a set of seven scoring rules in reducing the variability of the marks of 26 teachers on an arithmetic paper.

313. PARKER, E. P., "A Few Suggestions for Informal Testing in Geography," *Elementary School Journal*, Vol. 23, pp. 444-447, 1923.

314. PATERSON, D. G., "Improving the Examination Function in Teaching," *Bulletin of the University of Minnesota*, Vol. 26, No. 31, pp. 47-56, 1923.

315. PERRY, W. M., "Measurement and Analysis of Student Achievement in a Beginning Course in Educational Psychology," *Education*, Vol. 48, pp. 12-23, 1927.

316. POWERS, S. R., "Objective Measurement in General Science," *Teachers College Record*, Vol. 29, pp. 345-349, 1928.

317. PRATT, H. G., "Proper Use of Educational Measurements," *Journal of Educational Method*, Vol. 7, pp. 204-209, 1928.

318. PRESSEY, S. L., "A Simple Apparatus which Gives Tests and Scores and Teaches," *School and Society*, Vol. 13, pp. 373-376, 1926.

319. RAKESTRAW, N. W., "Objective Examinations in Chemistry," *Report of the New England Association of Chemistry Teachers*, Vol. 28, No. 5, pp. 133-143, 1927.
A very thoughtful discussion illustrated by unusually good concrete examples of objective test items for chemistry.

320. REEVE, W. D., "New-Type Tests in Teaching Mathematics," *Teachers College Record*, Vol. 29, pp. 693-703, 1928.
Rules for the construction of new-type tests together with the criteria of good examinations. Several excellent test exercises are given.

321. REMMERS, H. H., "The Relation Between Students' Marks and Students' Attitudes Toward Instructors," *School and Society*, Vol. 28, pp. 759-760, 1928.

322. RICH, S. G., "Use of Standardized and Partially Standardized Tests in a Normal School," *School Science and Mathematics*, Vol. 23, pp. 539, 542, 1923.

323. RICHARDS, O. W., "Test Construction in Less Standardized Subjects Illustrated by the Richards Biology Test," *School Science and Mathematics*, Vol. 27, pp. 22-27, 1927.

324. ROBACK, A. A., "Subjective Tests vs. Objective Tests," *Journal of Educational Psychology*, Vol. 12, pp. 439-444, 1921.

325. ROBINSON, E. S., "The Analysis of Trade Ability," *Journal of Applied Psychology*, Vol. 3, pp. 352-357, 1919.

326. ROGERS, D. C., "The New Type or Objective Examination," *Chicago Principals' Club Reporter*, Vol. 15, No. 4, pp. 3-6, 1925.

327. RUML, B., "The Extension of Selection Tests to Industry," *Annals of the American Academy of Political Science*, Vol. 81, No. 170, pp. 38-46, 1919.

328. SCHUTTE, T. H., "Is There Value in the Final Examination?", *Journal of Educational Research*, Vol. 12, pp. 204-213, 1925.

329. SHULSON, V. AND CRAWFORD, C. C., "Experimental Comparison of True-False and Completion Tests," *Journal of Educational Psychology*, Vol. 19, pp. 580-583, 1928.
The completion test lacks objectivity but is as valid as the true-false. The latter is easier, if not corrected for chance, but more difficult if scored right-minus-wrong.

330. SHERMAN, J. H., "Is the Examination Worth Retaining?" *School and Society*, Vol. 27, pp. 694-696, 1928.
Finds the examination so worth while that he recommends all grading upon such a basis.

331. SMITH, D. E., "On Improving Algebra Tests," *Teachers College Record*, Vol. 24, pp. 87-94, 1923.

332. STENQUIST, J. L., "Measurement of Mechanical Ability," Teachers College, Columbia University, *Contributions to Education*, No. 130, 1923.

333. STOCKS, E. H., "The Use of the True-False Examination at Smith College," *School and Society*, Vol. 22, p. 655, 1925.

334. STORMZAND, M. J., *American History Teaching and Testing*, New York, The Macmillan Company, 1926. 181 pp.
Chapters I, II, III, XXIV, and XXV deal with examinations, marks, new-type tests, etc., with special reference to instruction.

335. SYMONDS, P. M., "Needed Research in the Field of Measurement in Secondary Education," *Journal of Educational Research*, Vol. 16, pp. 119-127, 1927.
An excellent summary of the high points of the work done to date in high-school testing with several lists of needed researches, eighty such being suggested.

336. TELFORD, F. (Unsigned), "Methods of Selecting Employees to Fill High Grade Positions in Public Service," *Public Personnel Studies*, Vol. 2, No. 2, pp. 13-58, 1924.
The best statement in print of the values of and needs for objective tests for public employees. Many illustrative selection techniques (including sample tests) are given.

337. TELFORD, F., "The Work of the Board of Examiners of the New York City Board of Education," *Public Personnel Studies*, Vol. 2, No. 9, pp. 268-287, 1924.
Shows the extent to which this Board has made use of short-answer tests in selecting teachers.

338. THOMA, W. M., "Committee Marking of Examinations," *New York Bulletin of High Points*, Vol. 6, pp. 26-28, 1924.
Holds that committee marking is not sufficiently better than individual marking to justify its use.

339. THORNDIKE, E. L., "The Nature, Purposes, and General Method of Measurements of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education*, Part II, 1918.

340. TOOPS, H. A., "What Are We Failing to Measure?", *Journal of Educational Research*, Vol. 13, pp. 118-129, 1926.

341. TRYON, R. M., "Standard and New-Type Tests in Social Studies," *Historical Outlook*, Vol. 17, pp. 172-178, 1927.

342. VAN BUSKIRK, L., "Measuring Results of Physical Education," *Journal of Educational Method*, Vol. 7, pp. 221-229, 1928.

343. WAIT, W. T., "Objective Measurements of the Results of Solid Geometry," *School Science and Mathematics*, Vol. 17, pp. 969 ff, 1927.

344. WALKER, H. M., "Certain Mathematical Questions Suggested by the True-False Test," *American Mathematical Monthly*, Vol. 34, No. 10, pp. 503-515, 1927.
An excellent but highly mathematical consideration of the elements of probability in the true-false test.

345. WEISS, A. P., "On Methods of Mental Measurement Especially in School and College," *Journal of Educational Psychology*, Vol. 2, pp. 555-563, 1911.
So far as the present writer can determine, this is the first published study of the completion test. Many of the suggested studies have been carried out by later writers.

346. WEIDEMANN, C. C. AND WOOD, B. D., "A Survey of College Examinations," Bureau of Collegiate Research, Columbia University, April 1, 1926.

347. WELD, L. D., "A Standard of Interpretation of Numerical Grades," *School Review*, Vol. 25, pp. 412-421, 1917.

348. WHITTEN, C. W., "Report on Standardizing Teachers' Marks," *Sixth Yearbook of the National Association of Secondary School Principals*, 1922, Menasha, Wisconsin, The George Banta Publishing Company, pp. 183-202.
Recommends the abandonment of percentage numerical grades in favor of an A, B, C, D, and E system.

349. WILKINS, L. A., "Suggestions as to the Formulation of Questions in Standardized Examinations in Modern Languages," *New York Bulletin of High Points*, Vol. 5, No. 4, pp. 27-35, 1923.

350. WOOD, B. D., "Measurement of Law School Work," *Columbia Law Review*, Vol. 24, pp. 224-265, 1924, and Vol. 25, pp. 1-16, 1925.

351. WOOD, B. D., "New Type Examinations in the College of Physicians and Surgeons," *Journal of Personnel Research*, October-November, 1926, pp. 227-234 and pp. 277-283.
References 350 and 351 are accounts of pioneer work in applying the new-type examination in college classes in professional schools. Of especial interest to college teachers.

IX. SELECTED TEXTBOOKS ON EDUCATIONAL MEASUREMENT

352. BROOKS, S. S., *Improving Schools by Standardized Tests*, Boston, Houghton Mifflin Company, 1922. 273 pp.
A very elementary but practical discussion of the values of educational tests.

353. CORNING, H. M., *After Testing, What?* Chicago, Scott, Foresman and Company, 1926. 213 pp.
An account of the evolution and uses of a standard testing program in a small city school.

354. FENTON, N. AND WORCESTER, D. A., *An Introduction to Educational Measurements*, Boston, Ginn and Company, 1928. 149 pp.
A brief and very elementary treatment.

355. GILLILAND, A. R. AND JORDAN, R. H., *Educational Measurements and the Classroom Teacher*, New York, The Century Company, 1924. 265 pp.
One of the first treatments to bring the subject up to date at the time of publication.

356. GREGORY, C. A., *Fundamentals of Educational Measurement*, New York, D. Appleton and Company, 1922. 283 pp.
Valuable historical discussions and a general treatment of the idea of measurement.

357. GREENE, H. A., AND JORGENSEN, A. N., *Use and Interpretation of Educational Tests*, New York, Longmans, Green and Co., 1929, 411 pp.
A carefully prepared handbook designed for the use of the classroom teacher.

358. HULL, C. L., *Aptitude Testing*, Yonkers, The World Book Company, 1928. 535 pp.
One of the few scholarly treatments in the literature on educational measurement. Although the special emphasis is on the measurement of aptitudes, this volume should be mastered by all thorough students of testing.

359. KELLEY, T. L., *The Interpretation of Educational Measurements*, Yonkers, The World Book Company, 1927. 363 pp.
The best critical presentation of the theory and practice of educational test construction. Two especially valuable features are: (1) the evaluations of hundreds of educational tests, thus laying a basis for critical selection; and (2) the development of the statistical techniques related to educational measurement.

360. MCCALL, W. A., *How to Measure in Education*, New York, The Macmillan Company, 1922. 416 pp.
An early treatment of the theory and practice of test construction and related statistical procedures.

361. MONROE, W. S., *The Theory of Educational Measurements*, Boston, Houghton Mifflin Company, 1923. 363 pp.
A very readable account of how standard tests are validated and standardized.

362. MONROE, W. S., DEVOSS, J. C. AND KELLY, F. J., *Educational Tests and Measurements*, Boston, Houghton Mifflin Company, Revised Edition, 1924. 521 pp.
A revision of an early and widely used text. The tests available in the several school subjects are described.

363. PAULU, E. M., *Diagnostic Testing and Remedial Teaching*, Boston, D. C. Heath and Company, 1924. 371 pp.
A treatment of the follow-up work after testing.

364. PRESSEY, S. L., *Introduction to the Use of Standard Tests*, Yonkers, The World Book Company, 1922. 263 pp.
A book for the classroom teacher just beginning the study of measurements.

365. RUCH, G. M. AND STODDARD, G. D., *Tests and Measurements in High School Instruction*, Yonkers, The World Book Company, 1927. 381 pp.
A treatment of high-school tests containing a great many determinations of reliability and validity not previously available. Sections are provided on objective tests, test construction, and criteria for the selection of tests.

366. SMITH, H. L. AND WRIGHT, W. W., "Second Revision of the Bibliography of Educational Measurements," *Bulletin of the School of Education*, Indiana University, Vol. 4, No. 2, November, 1927. 251 pp.
The most complete and useful bibliography of educational tests which has appeared to date. Approximately one thousand titles are listed, together with short descriptions of each test.

367. SMITH, H. L. AND WRIGHT, W. W., *Tests and Measurements*, New York, Silver, Burdett, and Company, 1928. 540 pp.
Perhaps the most modern and best balanced treatment of the subject available. The authors have shown more tendency to depart from tradition than almost any other writers.

368. SYMONDS, P. M., *Measurement in Secondary Education*, New York, The Macmillan Company, 1927. 588 pp.
The most comprehensive and extensive treatment of high-school tests available.

369. TRABUE, M. R., *Measuring Results in Education*, New York, The American Book Company, 1924. 492 pp.
An extended treatment with special reference to the uses of test results in bettering classroom instruction.

370. VAN WAGENEN, M. J., *Educational Diagnosis*, New York, The Macmillan Co., 1926. 276 pp.
As the title suggests, the book deals with actual problems of measuring and classifying pupils, together with corrective work arising from the application of tests.

371. WILSON, G. M. AND HOKE, K. J., *How to Measure*, New York, The Macmillan Company, Revised Edition, 1928. 597 pp.
One of the better treatments of the subject. The authors have succeeded in keeping the needs of the teacher in mind better than many other writers.

372. WOOD, B. D., *Measurement in Higher Education*, Yonkers, 1923. 337 pp. (See Reference 23.)
No college teacher can afford to be ignorant of the pioneer investigations reported here on educational and mental testing in college instruction.

X. SELECTED REFERENCES ON STATISTICAL METHODS

373. GARRETT, H. E., *Statistics in Psychology and Education*, New York, Longmans, Green and Company, 1926. 331 pp.
On the whole, this volume is about as useful to the beginner as any existing treatment. There are especially strong treatments of measures of reliability and partial correlation. The needs of educators have been kept well in mind.

374. HOLZINGER, K. J., *Statistical Methods for Students of Education*, Boston, Ginn and Co., 1928. 372 pp.
A well-written and authoritative treatment, somewhat more advanced than Garrett, but within the ability of any student with a good working knowledge of algebra.

375. KELLEY, T. L., *Statistical Method*, New York, The Macmillan Company, 1923. 390 pp.
An advanced treatment of the field of statistics. The basic theorems of test validation and reliabilities are derived and applied. Some knowledge of higher mathematics will be needed to follow parts of the discussion.

376. OTIS, A. S., *Statistical Method in Educational Measurement*, Yonkers, The World Book Company, 1925. 337 pp.
A fairly elementary discussion of the handling and interpretation of test scores. Since this volume has been written with the needs of testers in mind, it is the best single reference for the test maker.

377. RUGG, H. O., *A Primer of Graphics and Statistics*, Boston, Houghton Mifflin Company, 1925. 142 pp.
A beginner's book on graphs and the more elementary statistical measures.

NOTE: Since this bibliography was prepared, Odell has published a fully annotated bibliography of 300 classified titles, as Bulletin No. 43 of the Bureau of Educational Research, University of Illinois, January 15, 1929.

INDEX

Abbott, A., 455
Accuracy of individual scores, 428-433
Achtenhagen, O., 464
Advantages of objective examinations, 112-120
Agard, A. F., 248-254
Agriculture examination of State of Wyoming, 224-229
Alameda High School literature test, 249-254
Alienation, coefficient of, 440-441; table of, 440
American history test of the Rochester schools, 229-240
American Library Association examination, 241-248
Analogies tests, examples of, 202
Answer keys or stencils, 172-184
Arithmetic mean, calculation of, 405-412
Army Alpha, studies on corrections for chance in, 333
Arnold, H. L., 459
Arrangement of items in order of difficulty, 34-36
Ashbaugh, E. J., 449
Asker, W., 324-327, 459
Attitudes of students toward examinations, 130-137
Average, calculation of, 405-412

Ballard, P. B., 359-360, 447, 459
Banker, H. J., 449
Bardy, J., 136, 452, 455
Barthelmess, H. M., 322-324, 459
Batson, W. H., 452
Bawden, W. T., 464
Berg, G., 455
Bernstein, L., 455
Best-answer tests, examples of, 199-200
Bingham, W. V., 464
Binomial theorem applied to test scores, 328-331
Bird, C., 464
Bliss, D. C., 449
Bluffing, in essay examinations, 108-110, 119; lack of in objective tests, 119
Blumer, G., 452, 464
Bolton, F. E., 85-89, 449
Boring, E. G., 459
Boyd, W., 452
Branom, M. E., 464
Breeze, R. E., 449
Briggs, T. H., 455, 456
Brinkley, S. G., 113-114, 116-118, 136-137, 281-283, 290, 306, 308, 309, 313, 346-347, 447, 453, 455, 459
Brooks, S. S., 469
Brown, C. M., 464
Buckner, C. A., 134, 455, 464
Building of objective tests, steps in, 149-187
Bureau of Public Personnel Administration, tests of, 456-457
Bursch, J. F., 447
Burton, W. H., 459
Butler, W. F., 454

Caldwell, O. W., 3, 450, 464
Camp, F. S., 449
Carlson, P. A., 457
Carpenter, M. F., 248
Carter, R. E., 448
Cattell, J. McK., 450
Chance, correction for, 318-357; formula for the allowance for, 185-187
Chapman, J. C., 457, 459
Charles, J. W., 289-290, 301-302, 306-316, 462
Charters, W. W., 464
Chassell, C. F. and E. B., 457
Chenoweth, L. E., 214
Cheydleur, F. D., 464
Christensen, A. M., 353-354, 460
Christensen, E. O., 458
Classification of objective tests, 188-190
Clatworthy, L. M., 241

Coefficient of alienation, 440-441; table of, 440
Coefficient of correlation, 412-420
Coefficient of reliability, definition of, 89-91; of new-type or objective tests, 291-306; of standard tests, 140-144; statistical calculation of, 412-420; statistical interpretation of, 433-441; summary of for traditional examinations, 106-108
College Entrance Examination Board, examinations of, 82-85, 464
Colvin, S. S., 143, 450
Comparable or equivalent forms, definition of, 66-67; making of, 164-166; values of, 67-69
Comparative difficulties of new-type or objective tests 314-317
Comparative working times for objective tests, 306-313
Completion tests, advantages and limitations, 271-272; examples of, 24, 192-193; rules for the construction of, 272-274
Computation tests, examples of, 204
Conneau, A., 188-189
Construction of objective tests, steps in, 149-187
Construction tests, examples of, 204
Cooley, A. M., 457
Cooprider, J. L., 457
Corning, G. B., 450
Corning, H. M., 469
Correction-of-error tests, examples of, 206
Corrections for attenuation, 285
Corrections for chance or guessing in multiple-response tests, formula for, 185-187; theoretical and experimental considerations, 318-357
Correlation, calculation of explained, 412-420
Courtis, S. A., 3, 21, 450, 464
Crathorne, A. R., 328
Crawford, C. C., 303, 453, 460, 467
Criteria of good tests or examinations, 27-69
Curtis, F. D., 354-355, 461
Cushman, C. L., 365

Dadourian, H. M., 464
Dalman, M. A., 457
Deduction tests, example of, 209
DeGraff, M. H., 116-117, 283-287, 290, 298-301, 306-313, 315-316, 336-345, 462
DeVoss, J. C., 142, 470
Dickinson, Z. C., 450, 464
Difficulties of objective or new-type tests, studies of, 314-317
Difficulty, arranging test items in order of, 163, 166; by judgment, 36
Dispersion, calculation of measures of, 422-428
Dolch, E. W., 458
Douglass, H. R., 460, 465
Doyle, L., 465
Drafting of objective test items, 153-159
DuBreuil, A. J., 457, 460
Duplicate or equivalent forms, definition of, 66-67; making of, 164-166; values of, 67-69
Durrell, D. D., 365

"Ear marks" of a good test or examination, 27-69
Ease of administration as a criterion of a good test, 63-64
Ease of scoring as a criterion of a good test, 65
Eaton, M. P., 457
Economy of scoring of objective tests, 118-119
Economy of testing, lack of in essay examinations, 108
Elimination of examinations, 8-10
Elliott, E. C., 77-78, 81, 452
Ellis, R. S., 386-388, 447, 465
Elston, B., 454
English examinations, 248-254
Equivalent or duplicate forms, definition of, 66-67; making of, 164-166; values of, 67-69
Error of estimate, 437-441
Error of measurement, 45-48, 428-433
Essay or traditional examination, definition of, 18; objections to, 70-111
Evans, M. E., 457
Everett, S., 229
Ewing, D. H., 466

Examinations, criteria of, 27-69; functions of, 10-18; history of, 3-8; principal types, 18-26; retention or elimination of, 8-10; students' attitude toward, 130-137
Experimental studies on new-type or objective examinations, 281-317
Extensive sampling, 56-59, 114-115

Farwell, H. W., 457
Fauber, O. W., 102-106
Feingold, G. A., 450
Fenton, N., 465, 469
Filer, H. A., 457
Finkelstein, I. E., 450
Fiske, T. S., 453
Flack, W. S., 465
Foote, M., 465
Foran, T. G., 460
Foster, R. R., 348-353, 460
Foster, W. T., 450
French, H. P., 450
Fritz, M. F., 365, 460
Functions of examinations, 10-18

Gallagher, E. D., 132-133
Galton, F., 68
Garrett, H. E., 471
Gates, A. I., 453, 458, 460
Gathany, J. M., 465
Gaw, E. A., 450
Gerry, H. L., 465
Gifford, C., 48
Giles, J. T., 460
Gilliland, A. R., 469
Glenn, E. R., 453, 465
Good, H. G., 465
Goodale, M., 458
Gordon, H. C., 458
Gordon, W. C., 97-98
Grading systems, 369-402
Gray, C. T., 450
Gray, W. S., 454
Greene, C. E., 447
Greene, H. A., 353-354, 460, 469
Gregory, C. A., 469
Griffin, H. D., 460
Guessing, corrections for, 318-357; effects of instructions on, 339-343; in objective tests, 120
Gundlach, R., 460

Hahn, H. H., 320-324, 460
Hammond, E. L., 460
Hance, R. T., 465
Hannig, W. A., 453
Hatch, R. W., 465
Hawkes, H., 465
Healy, H. L., 214
Heffernan, H., 454
Hemstreet, A. E., 465
Hendrickson, C. E., 71, 75, 450
Herring, J. P., 454
Heterogeneity or range of talent, 441-445
History of examinations, 3-8
Hoke, K. J., 471
Holzinger, K. J., 324-328, 460, 471
Hopkins, L. T., 447
Horn, E., 29
How to construct objective examinations, 149-187
Hughes, R. O., 134-135, 455, 464
Hull, C. L., 470
Hulten, C. E., 450

Identification tests, examples of, 205
Illustrative tests and test items, 191-264
Informal examinations, definition of, 9
Inglis, A., 465
Instructions for objective tests, 166-172
Intensive sampling, 56-59
Irving, H., 224, 454

James, B. J., 453
James, H. W., 450, 451, 461
Jennison, H. M., 465
Johnson, F. W., 70, 75, 451, 465
Jordan, R. H., 469
Jorgensen, A. N., 469
Judgments of item difficulties, 36
Justification for examinations, 8-10

Kansas Scholarship Contest, 458
Karwoski, T. R., 458
Kelley, T. L., 61, 142, 346, 439, 440, 470, 471
Kelly, F. J., 70, 73, 101-102, 103, 142, 451, 470
Kent, S. T., 241

Kepner, T. P., 466
Kern County, California, examinations, 214-224
Kimball, E. P., 466
Kimmell, W. C., 466
Kimmell, W. G., 454
Kinder, J. S., 136, 137, 455
Klise, N. W., 455
Knight, F. B., 453
Kohs, S. C., 324-328, 461, 462
Kolstoe, S. O., 136, 137, 455
Kwalwasser, J., 254-262
Kwalwasser-Ruch Test of Musical Ability, 254-262

Lagergren, A. C., 241
Laird, D. G., 453, 461
Langlie, T. A., 288-290, 334-335, 462
Language training, examinations for, 13-16; neglect of in objective examinations, 120
Lathrop, H. O., 458
Laycock, S. R., 458
Lee, B., 461
Lehmann, H. C., 461, 465
Library examinations, 241-248
Limitations of objective examinations, 120-129
Literature, examinations on, 249-254
Lohr, V. C., 453, 458
Louterbach, C. E., 451
Lowell, A. L., 466

MacPhail, A. H., 466
Making of objective examinations, steps in, 149-187
Maloney, E., 454
Mann, Horace, views on examinations, 3-8, 18
Map location tests, example of, 208
Marks and marking systems, a matter of definition, 376-378; classification of, 369-370; normal curve in assigning marks, 374-392; percentage marking systems, 370-374; per cents vs. ranks in marking, 392-402; reliability of teachers' marks, 89-111; variations in teachers' marks, 70-89
Marsh, W. R., 466
Matching tests, advantages and limitations of, 276-277; examples of, 200-202; rules for construction of, 277-278
Mathews, C. O., 344, 366-367, 461
Maupin, N., 303-306
May, M. A., 136, 137, 455, 458, 461
McAfee, L. O., 461
McCall, W. A., 168, 321, 461, 466, 470
McClusky, H. Y., 354-355, 458, 461
McGregor, J. B., 91-96
McLeod, B., 224, 454
Mead, A. R., 466
Mean, calculation of, 405-412
Measurement, as ranking, 392-402; complete vs. incomplete, 48-56; examinations as, 16-18
Meltzer, H. M., 447
Melvin, A. G., 458
Memory vs. thought questions, 120-126
Mersereau, E. B., 451
Meyer, G., 447
Meyer, H. W., 366
Meyer, M., 374, 451
Miles, W. R., 451
Miller, G. F., 448, 461
Miller, W. S., 458
Mimeographing of examinations, 126
Miscellaneous tests, 263-264
"Missouri system" of marking, 374
Mitchell, S. B., 241
Monroe, W. S., 91, 97, 106-107, 142-143, 144, 448, 461, 470
Morley, E. E., 466
Motivation, examinations for, 10-12
Moyer, F. E., 458, 466
Muenzinger, K. F., 461
Multiple-response or multiple-choice tests, advantage and limitations of, 274-275; examples of, 24, 198-200; experimental studies on scoring of, 345-353; rules for the construction of, 275-276
Murdoch, J. R., 303-306
Music tests, 255-262

"Natural" vs. "unnatural" forms of questioning, 126-128
Negative suggestion effects in true-false tests, 358-368

New-type or objective tests, advantages of, 112-120; building of, 149-187; definition of, 9; experimental studies of, 281-317; limitations of, 120-129; samples of, 191-212, 213-264
New York Regents' Examinations, 97-99
Normal curve in marking, discussion of, 374-388; limitations of, 381-383; when classes are sectioned upon a basis of ability, 388-392; when classes are small, 383-388
Norms or standards, importance of, 66

Objective or new-type tests or examinations, advantages of, 112-120; building of, 149-187; definition of, 9, 23-26; experimental studies of, 281-317; limitations of, 120-129; relative values of, 138-146; samples and illustrations of, 191-264
Objective tests compared with standard tests, 138-146
Objectivity in relation to reliability, 42-45
Odell, C. W., 47-51, 322-324, 369, 448, 451, 458, 461
Ogan, R. W., 466
Opdyke, J. R., 466
Oral examinations, 18-19
Order of difficulty for test items 34-36, 163, 166
Orleans, J. S., 214, 448, 458
O'Rourke, L. J., 457
Osburn, W. J., 113
Otis, A. S., 139-141, 432, 471
Ovitz, D. V., 241
Ozanne, C. E., 455

Parker, E. P., 466
Paterson, D. G., 266, 288-289, 290 334-335, 448, 453, 462, 466
Paulu, E. M., 470
"Pedagogical" vs. "unnatural" questions, 126-128
Percentage grading systems, 370-374
Per cents vs. ranks in grading, 397-402
Perry W. M., 466
Powers, S. R., 466
Pratt, H. G., 466
Pressey, S. L., 143, 466, 470
Probability, theory of applied to tests, 320-331
Probable error of test score, 428-433
Public Personnel Studies, tests published in, 456-457

Rakestraw, N. W., 467
Range of talent or heterogeneity, 441-445
Ranking, measurement as, 392-402
Ranks vs. per cents in marking, 397-402
Rating of test items for difficulty, 163-164
Raynaldo, D. A., 303, 453, 460
Rearrangement tests, examples of, 202-3
Recall tests definition of, 23; examples of, 191-194; rules for the construction of, 269-271
Recognition vs. recall in examinations, 128-129
Redundancy tests, examples of, 206-208
Reeve, W. D., 467
Reeves, G., 457
Regrading of essay examinations by different teachers, 77-106
Regression effects in test scores, 68
Regression equations, use of in prediction, 435-441
Reliability and validity compared, 59-62; coefficient of, defined, 89-91; definition of, 40-41; effects of corrections for chance on, 334-335; interpretation of, 433-441; methods of insuring, 42; of objective or new-type examinations, 291-306; of standard tests, 140-144; of state eighth-grade examinations, 91-96; of teachers' marks, 70-107; of traditional or essay examinations, 106-108; relation of objectivity to, 42-45; relation of sampling to, 45-49; statistical determination of, 412-420
Remmers, H. H., 360-361, 462, 467
Remmers, H. H. and E. M., 360-361, 462
Rice, G. A., 25, 188, 213, 248
Rice, J. M., 6, 12, 21

Rich, S. G., 451, 462, 467
Richards, O. W., 324-328, 462, 467
Rietz, H. L., 328
Right-minus-wrong method of scoring, 331-357
Right-wrong tests, examples of, 196-197
Riley, E., 229
Roback, A. A., 453, 467
Roberts, H. M., 361-364, 462
Robinson, E. S., 467
Rochester schools, American history examination, 229-240
Rogers, D. C., 467
Ruch, G. M., 17, 25, 28, 30, 56, 66, 75, 78-82, 91-96, 102-106, 117, 132-133, 142, 188, 213, 248, 255-262, 283-287, 290, 292-296, 298-301, 303-306, 307-317, 334-357, 361-364, 432, 448, 451, 454, 460, 462, 470
Rugg, H. O., 451, 471
Ruml, B., 467
Russell, C., 449
Rutledge, R. E., 302-303, 366, 462

Samples of objective test items and examinations, 191-264
Sampling, theory of, applied to examinations, 30, 52-59, 114-115; relation to reliability, 45-49
Sandon, F., 452
Sanford, V., 453, 458
Schryoch, R. H., 453
Schutte, T. H., 454, 467
Scoring or answer keys and stencils, 172-184
Scoring tests, rules for, 184-187
Sealy, J. A., 214, 448
Sectioned classes, marking of, 388-392
Sharp, L. A., 452
Sherman, J. H., 467
Short-answer tests, definition of, 9; examples of, 194
Shriner, W. O., 452
Shulson, V., 467
Siegrist, S., 447
Simple-recall tests, advantages and limitations of, 269; examples of, 191-192; rules for constructing, 270-271
Skinner, A. W., 453
Smith, D. E., 467
Smith, H. L., 470
Somers, G. T., 131-134, 455
Souders, L. B., 91, 448
Spearman, C., 285
Spearman-Brown formula, 294, 296, 418, 420-422
"Specific determiners," 319
Spence, R. B., 452
Spencer, P. L., 454, 460
Stack, H. J., 458
Standard deviation, calculation of, 422-428
Standard error of estimate, 437-441
Standard tests, definition of, 21-23; relative values of, compared with unstandardized (or objective) tests, 138-146
Standards, values of examinations for maintaining, 12-13
Starch, D., 77-78, 81, 452
State eighth-grade examinations, studies of, 91-96
State of Wyoming, examination in agriculture of, 225-229
Statistical methods related to measurement, 405-445
Stencils or answer keys, 172-184
Stenquist, J. L., 467
Stock, E. H., 467
Stoddard, G. D., 17, 28, 30, 66, 117, 142, 292-296, 306, 307, 308, 309, 314, 315, 334-335, 448, 462, 470
Stone, C. W., 21
Stormzand, M. J., 468
Strang, R., 458
Strickland, V. L., 449, 458
Students' attitudes toward examinations, 130-137
Subjectivity of marking, 21, 77-106; reduction of by means of scoring rules, 101-106
Suggestion effects in true-false tests, 358-368
Summarizing test scores, statistical methods of, 405-412
Sutherland, A. H., 454
Symonds, P. M., 17, 99, 142, 386-388, 449, 462, 468, 470

"Table of Specifications," 150-152
Talbott, E. O., 52-55

Teachers' marks, investigations of, 70-106
Telford, F., 456-457, 463, 468
Test items, order of difficulty of, 34-36, 163, 166; rules for drafting, 265-278
Test publishers, list of, 264
Test scores, statistical methods of summarizing, 405-412
Tharp, J. B., 454
Thoma, W. M., 468
Thomson, G. H., 113
Thorndike, E. L., 21, 29, 452, 459, 468
Thorndike-McCall Reading Scales, 57
Thought vs. memory questions, 120-126
Thurstone, L. L., 328, 345-346, 349, 463
Tidyman, W. F., 452
Toops, H. A., 116, 291-292, 306, 307, 308, 309, 313, 314, 449, 457
Trabue, M. R., 471
Traditional or essay examinations, definition of, 18; discussion of, 20-21; objections to, 70-111
Translation tests, examples of, 209-210
True-false tests, 24; advantages and limitations, 265; as two-response tests, 343-345; examples of, 194-198; rules for the construction of, 265-269; variations of, 353-355
Tryon, R. M., 468

"Unnatural" vs. "pedagogical" questions, 126-128
Unreliability, due to limited sampling, 45-59; due to subjectivity, 42-45; of teachers' marks, 70-106
U. S. War Department, 463

Validation, methods of, 28-39; of test items, 31
Validity and reliability contrasted, 59-62; defined and discussed, 27-40; effects of corrections for chance on, 335-339; of new-type tests, 281-290
Van Buskirk, L., 468
Van Wagenen, M. J., 471
Variation, calculation of measures of, 422-428

Wait, W. T., 468
Walker, H. M., 468
Waples, D., 455
Webb, H. A., 459
Weidemann, C. C., 33, 266, 319, 449, 463, 469
Weinland, J. D., 463
Weiss, A. P., 461
Weld, L. D., 469
West, P. V., 324-328, 463
Whitten, C. W., 469
Wigmore, J. H., 463
Wilkins, L. A., 469
Wilson, G. M., 29, 471
Wilson, H. E., 463
Wood, B. D., 9, 82-85, 97-100, 106, 116, 124-126, 168, 287-288, 290, 296-298, 335-339, 449, 454, 463, 469, 471
Wood, E. P., 338-339, 348-353, 463
Worcester, D. A., 463, 469
Working time required for objective or new-type tests, comparative studies of, 306-313
Wright, H. C., 459
Wright, W. W., 459, 470
Wyoming state examination in agriculture, 225-229

Yerkes, R. M., 333, 463
Yes-No tests, examples of, 195-196

This book may be kept
FOURTEEN DAYS
A fine will be charged for each
day the book is kept overtime.

OCT 13 1983
OCT 25 1983
FEB. 1 0 1986

Demco 38-293

THE MARY E. BIVINS LIBRARY
CINCINNATI BIBLE COLLEGE & SEM. LIBRARY
371.271 R899 main
Ruch, Giles Murrel/The objective or new-
3 4320 00002 9399

P. W. La

371.271
R899o 27203

Ruc
T
Exa

THE CINCINNATI BIBLE
SEMINARY LIBRARY
A. No. 27203
D. D. No. 371.271
R899o